Methods of Modelling the Mental Representation of Individuals Derived From Descriptions in Text

Joseph Patrick Levy

PhD

University of Edinburgh

1990

I declare that this thesis has been composed by myself. Unless otherwise stated, the work described is my own. My contribution to the teamwork involved is described in Section 1.7.

Acknowledgements

I would like to thank Keith Stenning for his help, guidance, support and pints of Guinness. I would also like to thank all the members of the "text group" (even the silly ones). I am grateful for the computational assistance given by Robert Dale and Andrew Zissermann.

This thesis would not have been produced without the support and friendship of my family and friends.

Abstract

The thesis describes the methods that were used to gain a fuller understanding of how humans represent the attribution of properties to individuals. Despite its neglect in the literature, this *attribute binding problem* is shown to be non-trivial and fundamental to many aspects of cognition.

The development of an experimental paradigm designed to collect data pertinent to this problem is described. The paradigm (the *memory for individuals task*) collects reading time data and recall error data arising from the interpretation and recall of simple texts describing a number of individuals and their properties. A number of experiments are described that demonstrate the wide applicability of the paradigm.

The reading time data is interpreted as measuring the relative cognitive loads imposed at different points in the texts. The work gains its impetus from an observation (Stenning 1986) that the reading time for a sentence about an individual increases as the individual becomes more specified. A successful statistical model of the reading time data was constructed that decomposes the cognitive load applied during the interpretation of the text into separate components. These components reflect different aspects of the informational structure of the descriptions.

The representation built is probed by an analysis of the errors that subjects make. There are a striking number of multiple errors, many of which fall into clearly defined categories. The errors are shown to reflect the dependencies between different parts of the underlying representation. A second statistical model is built that can account for the different classes of errors that subjects make. The statistical procedure extracts a similarity metric from the data that is used to define a fragmented representation with considerable redundancy.

The statistical model of recall error is used to define a PDP network that can synthesise the fragments of the representation into a correct recall. The network model is also capable of resolving the inconsistencies that arise when the redundant representation is subjected to random noise. The inconsistency resolution process is shown to be a plausible mechanism for the making of errors in recall since the networks make the same categories of errors that occur in the human data

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter contains some initial discussion on the nature of the 'attribute binding problem' — the problem of representing the relationship between an individual entity and its properties. The importance of content and context is discussed and a brief survey of relevant previous models of knowledge representation, most of which have not considered attribute binding to be a problem, is presented. The paradigm that was used for the work described in this thesis was first described in Stenning (1986). This work is summarised and an outline of the general paradigm that was developed from it, allowing the collection of both reading times and recall error data, is outlined. The introduction ends with summaries of the aims, structure and theoretical and methodological biases of the thesis.

1.1 The Attribute Binding Problem

Much of recent cognitive psychology, cognitive science and artificial intelligence has been concerned with ways in which knowledge can be represented. There have been complex debates on such issues as imagistic representation versus propositional representation (e.g. Pylyshyn 1973, Anderson 1978, Johnson-Laird 1983 Chapter 7), the representation of word meaning (e.g. Miller and Johnson-Laird 1976), and formal methods for implementing knowledge representation in computer programs (e.g. Brachman and Levesque 1985). From the breadth of this

debate the casual observer might be forgiven for assuming that some simple knowledge representation problems must already have been solved. There must surely be a complete understanding of how an individual object and its properties can be represented. Surely a computer database for a company payroll is a sufficient demonstration that this 'problem' has already been solved.

The following might be how a company payroll program stored a description of two employees, Smith and Jones:

```
(employee   (name john_smith)
            (address 246_high_street)
            (job clerk)
            (salary peanuts))

(employee   (name mary_jones)
            (address 1_the_grange)
            (job director)
            (salary too_much))
```

This is a convenient and efficient way of storing information for certain purposes. The database has a well ordered organisation and can be implemented in a purely *structural* way in terms of locations in computer memory and pointers to locations in memory (see Stenning and Levy 1988). The attribution of a property to an individual or object is achieved almost effortlessly by taking advantage of the structural primitives of the serial computer and its operating software.

The representation of similar information in human memory is likely to be very different however. I might remember John Smith as the man who has the same common name as a member of the shadow cabinet and who lives above the nice wine shop in town but can only afford to buy supermarket plonk. My representation for this individual is likely to be less rigidly ordered than the computer database and far more based on *content*. Somehow I am able to retrieve all kinds of relevant and irrelevant information from memory in a seemingly effortless and content addressable manner.

An example of a slightly harder problem in everyday life where the correct attribution of properties to individuals is involved might be a waiter trying to remember

that it was the man in the red shirt who ordered ice cream and an espresso while the man in the blue shirt had asked for ice cream followed by a cappuccino. Although only a small amount of information needs to be remembered and for only a relatively short time, a problem is posed by having to represent *which* property is attributed to *which* individual. For a computer a database solution like the one described above would suffice here, but it would seem, judging by the number of times my order in a restaurant has been confused with someone else's, that humans find this particular problem harder than computers do. Presumably the mechanisms that allow easy access to John Smith's particulars makes representing who ordered which kind of coffee difficult.

We have called the task of representing which properties have been attributed to which individuals the *attribute binding problem* (Stenning and Levy 1988). This thesis discusses the methods that we have developed to allow the investigation of human solutions to this problem. The next section briefly describes some of the evidence that supports the importance of content in human memory and comprehension in general, and the attribute binding problem in particular.

1.2 The Importance of Content and Context

The crucial importance of the effect of the interpreted content of material to be remembered on its recall is well established in the literature. A simple example is Miller's (1956) discussion of how material can be recoded or "chunked" on the basis of past knowledge. Bartlett (1932) put forward a theory of memory for contentful material based on Head's notion of the organisation of past sensory stimuli into 'schemata'. A schema is a flexible conceptual structure that serves to organise past experiences of a particular phenomenon. Bartlett's 'constructive character of remembering' depends on the assimilation of new material into existing schemata. This can only happen if the material to be remembered has been interpreted on the basis of its meaning or content.

Bransford and his colleagues have performed many 'demonstration experiments'

that show the necessity of having some context to allow access to relevant information in the knowledge-base while a fairly complex text is being read. For example, Bransford and Johnson (1972, 1973) designed their famous 'balloon text' so that subjects would require some sort of context for it to be interpreted. They found that the best comprehension was produced when a full pictorial context was shown to the subjects before they heard the text. The best recall was produced for conditions where subjects rated the texts most comprehensible i.e. those where there was sufficient context. The fact that showing the subjects the picture after they heard the text did not produce a good recall score suggests that the context is needed while the text is being comprehended and the representation of information from the text is dependent on this semantic interpretation. They also showed that the comprehension of a text can be greatly improved by giving an appropriate title that would allow comprehension to be placed in the context of stored knowledge.

Bransford et al. (1972) demonstrated that the comprehension of a text often involves some sort of inference based on general background knowledge. Subjects listened to one or the other of the following sentences:

(i) Three turtles rested *beside* a floating log, and a fish swam *beneath them.*

(ii) Three turtles rested *on* a floating log, and a fish swam *beneath them.*

After several sentences a recognition test was given, using sentences like:

(iii) Three turtles rested *beside* a floating log, and a fish swam *beneath it.*

(iv) Three turtles rested *on* a floating log, and a fish swam *beneath it.*

Subjects who heard (i) rejected both (iii) and (iv), while those who heard (ii) recognised (iv) and rejected (iii). The explanation that Bransford et al. gave was that (iv) was a plausible inference that followed from (ii) since a fairly large fish would be likely to pass under the turtles if it swam under the log they were resting on. However, (iii) does not follow from (ii) and neither (iii) nor (iv) follows from

(i).

Human memory, it seems, is based on a representation that is intimately linked with general knowledge. The richness of this representational form is what gives human memory its extraordinary capabilities — its seemingly unlimited capacity and efficient information access. Transparent and effortless access to background knowledge seems to be a feature of many cognitive tasks from the understanding of continuous speech (see Frauenfelder and Tyler 1987 and Lowe 1988 for some recent work) to anaphor resolution (e.g Sanford and Garrod 1981).

Bartlett and those like Bransford and his colleagues who have followed in his tradition have demonstrated the importance of the influence of past knowledge on the representation of new information. They did not, however, present any explicit representational form for human memory.

1.3 Stenning (1986)

This section discusses some empirical results from an experimental paradigm that allows the investigation of the construction of representations of simple descriptions. The materials used were contentful but easily manipulable. The paradigm formed the basis for the work reported in this thesis.

The first experiments to use what we now call the Memory for Individuals Task or MIT were reported in Stenning (1986). Stenning set out to study the constructive processes underlying the comprehension of simple texts. The texts used were carefully constructed so that their size and complexity could be controlled. The theoretical question behind the design of the experiments was how people construct a *model* of the information in a text — a set of objects mapped onto the set of predicates and relations of the text. This approach was developed in Stenning (1975, 1978, 1980) and bears strong affinities to some of the work of Johnson-Laird and his co-workers summarised in Johnson-Laird (1983). Johnson-Laird however, uses a model-like formalism as a representational medium whilst

No. of individuals	No. of adj. dimensions	No. of models
2	3	36
4	2	35
3	3	120
2	4	136

Table 1.1: The number of possible logical models for a given number of individuals and binary adjective dimensions

Stenning simply claims that a model must be represented.

Two studies were reported. The first set out to be a collection of base-line data to compare with the data from the second study. As it turned out, the 'base-line' data was interesting enough to spawn a series of experiments for our whole research group. The first study used various different types of determinate descriptions i.e. descriptions where the assignment of property to an individual was unambiguous at each sentence. The second study used texts where there was temporary indeterminacy — attribute binding had to be delayed.

The first study used four different kinds of description. They were distinguished by the number of individuals described and the number of different adjective dimensions used to classify the individuals. An adjective dimension was simply a contrasting pair of adjectives e.g. black versus white. Each of these different kinds of description has a different number of possible logical models for a given set of individuals and adjectives. The details of the different model structures are given in Table 1.1. As can be seen in the table, the four model structures can be divided into two pairs of structures whose members have roughly equal complexity in terms of number of possible logical models.

The experiment used two different *text modes* or temporal orderings of information. [1] The first mode is one where one individual is described and then the next one is described until all the individuals have been described. This is an *Individual by Individual* or IxI text, e.g:

[1]The definitions of terminology used in this chapter are the ones that will be used in the rest of the thesis and not always the same terms used in Stenning (1986).

There is is a square.
The square is white.
The square is large.
There is a circle.
The circle is black.
The circle is large.

The other text mode, *Property by Property* or PxP describes all the individuals along one dimension and then along the next dimension until the description is finished, e.g:

There is a square.
There is a circle.
The square is white.
The circle is black.
The square is large.
The circle is large.

Each description was prefixed by a setting that showed the dimensions on which the individuals would be classified and a sentence to tell the subject how many individuals would be described by the text. After the description the subject was asked two questions and prompted to recall the information in the text. These aspects of the task would have been as follows for the previous examples of descriptions:

Setting: Black/white, circle/square, large/small
Number statement: There are just two objects.
DESCRIPTION
First question: Is there a large black object?

Second question: Are there two large objects?
Recall prompt: What is there?

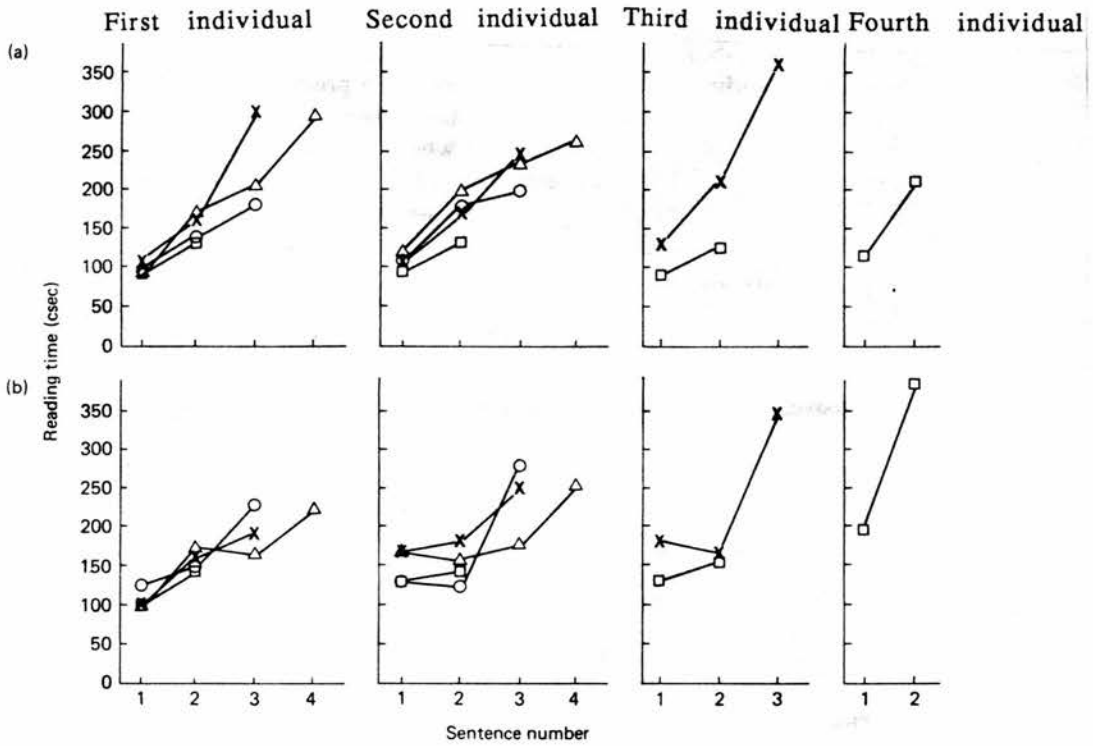
The vocabulary used to generate the descriptions came from four groups of six vocabulary pairs. A pair was chosen from as many groups as needed to make up a description. The groups corresponded roughly to magnitudes, textures, colours and shapes.

The descriptions were read sentence by sentence in a self-paced manner from a computer monitor. The reading time for each sentence was recorded. Reading time was chosen as the most direct way to measure the constructive processes going on during the building up of the representation of the descriptions.

Stenning's major observation in the first study was one that triggered many further experiments and formed the impetus for the work described in this thesis. He found that reading times increased with the successive specification of the properties of an individual. When a new individual was introduced reading times dropped. This effect was independent of the text mode — when the reading times of the sentences in a PxP text were plotted in terms of the properties attributed to each individual instead of in temporal order, the curves took exactly the same form as those of IxI texts — a gradual increase in reading times as an individual was described and then a speeding up as a new individual was introduced. The reading time results are summarised in Figure 1.1.

The fact that this effect took place for PxP texts where the description of different individuals was interleaved made it unlikely that the increase in reading times was the result of maintenance processes accumulating material before interpreting it (see Jarvella 1971, Baddeley 1986) since there would presumably have to be several parallel maintenance processes going on for this to be the case.

The form of the increasing reading times as an individual became more specified was called the *semantic ordinal effect* (SOE) by Stenning, Shepherd and Levy (1988) — 'semantic' because of the claim that the reading times are reflecting



Mean Sentence-Reading Times for determinate texts specifying four model sizes.
(a) Property-by-property (PxP) mode
(b) Individual-by-individual (IxI)mode

△ two individuals, four properties
○ two individuals, three properties
× three individuals, three properties
□ four individuals, two properties

Figure 1.1 Reading Times for Determinate Texts

constructive interpretational processes and 'ordinal' because they seem to depend on how many previous sentences there have been about the individual being described at any particular point.

The second study was one where the descriptions were temporarily indeterminate, forcing the process of attribute binding to be unpredictably delayed. The form of the experiment was similar to the first study. All the descriptions consisted of two individuals classified on three dimensions. There were four text forms, ranging from a control determinate form to one which had a period of indeterminacy of three sentences, e.g. for two objects, A and B:

1. There is a small black object. (A)
2. There is a small square. (?)
3. There is a small white object. (B)
4. There is a white square. (B).
5. There is a black circle. (A)

The attribution of property to individual that can be made after each sentence is denoted by either (A) or (B). The attribution of 'small square' at sentence 2 is delayed until sentence 5 where it can be inferred that the small square is white and that sentence 2 is describing object B.

The most important result of this study was that the load imposed by the indeterminacy of the descriptions and reflected in an increase of reading times did not occur at the onset of indeterminacy but was delayed until the point where the resolution of the indeterminacy could occur.

The material used in the second study flouts the normal conventions of expository text that ensure that descriptions are built up in a determinate fashion. By exposing the inferences that are required for attribute binding, it demonstrates that this process is indeed non-trivial and can be fruitfully probed by reading time techniques.

The next section reviews some of the major theories of human knowledge representation. Most of them ignore attribute binding and none of them can account for the semantic ordinal effect observed by Stenning.

1.4 Models of human knowledge representation

1.4.1 Propositional and semantic network theories

One of the first general knowledge representation formalisms used in artificial intelligence was the semantic network. A semantic network is a network of nodes that represents words or concepts associated by labelled links. The labels on the links express the semantic relation that holds between two nodes. Semantic networks have been used as the basis for many models of human knowledge representation (see Baddeley 1976 Ch. 13; Johnson-Laird et al. 1984). As well as specifying how meaning (or at least, some intensional relations between different concepts) can be expressed in network form, theories based on this kind of representation must also explain how the structure is built and how it is searched to enable the use of its stored information.

Quillian's (1968) theory is generally acknowledged to have been one of the earliest important contributions in this field. He proposed a model of lexical knowledge expressed as a semantic network. The meaning of a word is represented by a network of *type* nodes, representing concepts linked to *token* nodes which represent instances of concepts. Collins and Quillian (1969) attempted to test the psychological validity of the model. In the putative memory structures they used each node was linked to its *properties*. An example of a three-level hierarchy of concepts and instances is given in Figure 1.2. Each property is stored with the highest level concept to which it applies, allowing a certain economy of storage. To confirm that a canary has wings a path is traced through the network from 'Canary' to 'Bird' and then from 'Bird' to 'Has Wings'. Collins and Quillian showed that sentence verification time appeared to depend on the number of links through the network that had to be traced to verify the meaning of the sentence. There are

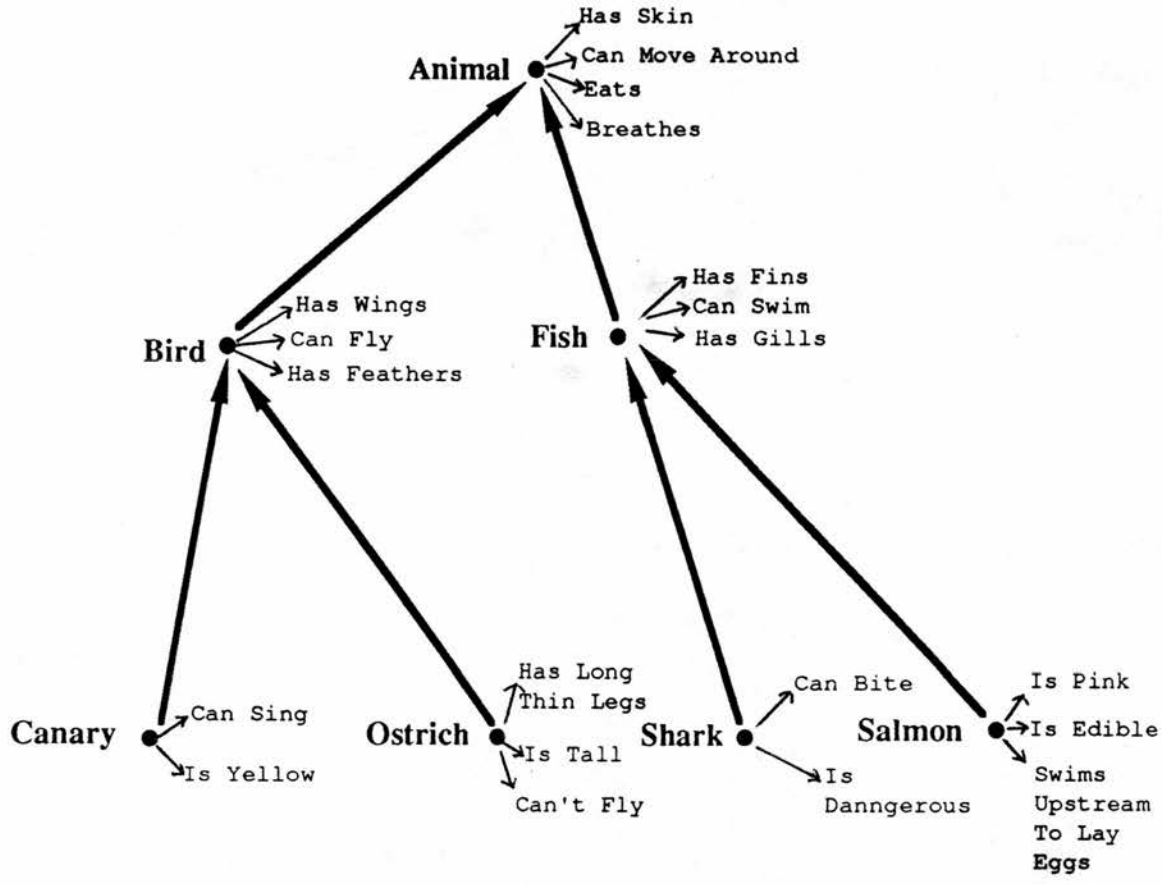


Figure 1.2: Example of a three-level conceptual hierarchy from Collins and Quillian (1969)

many problems with the model, allowing alternative explanations of their data (e.g Landauer and Meyer 1972; Conrad 1972) but what is of more importance in the context of the work to be presented in this thesis is the way in which this kind of model treats the attribute binding problem. Attribute binding is represented by a purely structural, primitive contentless link (e.g. between 'Canary' and 'Is Yellow'). These workers were concerned with the problems of representing intensional relations in a way that expressed the meaning of words and might allow the modelling of comprehension. The structure of computer hardware, the programming languages available at the time (see Johnson-Laird et al. 1972 p. 294) and perhaps the influence of predicate logic must have made this seem an entirely natural and unproblematic method of achieving attribute binding. What was ignored was the use of content in the representation of attribute binding, and the possibility that attribute binding might not be a cognitively trivial problem. It is not surprising, then, that these models cannot account for Stenning's reading time data. They provide no mechanism that can explain why attribute binding (reflected in reading time) should become more demanding as the description of an individual becomes more detailed. In this thesis we develop models that can account for Stenning's results. The use of content and general knowledge are essential elements of these models.

Perhaps one of the most ambitious research programs based on semantic networks is that of Anderson and his co-workers in the development of the HAM, ACT and ACT* models. Anderson and Bower's (1973) HAM model was a relatively simple model of associative memory incorporating a simple parser of English and mechanisms for retrieval. ACT (Anderson 1976) and ACT* (Anderson 1983) are more complex and include production systems (mechanisms for the application of *if ... then ...* rules) as well as semantic networks. Anderson (1983) has claimed that ACT* is a "theory of cognitive architecture" and a "unitary theory of mind". ACT* is an impressive achievement, encompassing a wide range of phenomena and data. However, like the many other theories employing semantic networks (e.g. Rumelhart et al. 1972; Glass and Holyoak 1975; Collins and Loftus 1975), ACT* achieves attribute binding merely by linking an individual and property with a link labelled with something like **'IS'**.

If semantic network theories had taken attribute binding seriously they might have been able to attempt to account for the semantic ordinal effect as some sort of search phenomenon. Collins and Quillian modelled verification time as an effect of the number of links that had to be traversed and Anderson has discussed the 'fan effect' where the spreading of activation through a network is made more diffuse the deeper or wider the path it traverses. These effects take place *after* the representation of attribute binding has taken place. There is nothing in the structure of a simple **IS** link that can account for why the semantic ordinal effect occurs *as* attribute binding occurs. The fact that these models cannot account for Stenning's data is a consequence of theoretical background and empirical methods that assumed attribute binding to be trivial. Because of this the semantic ordinal effect was not observed and so did not need to be accounted for.

Recent work in Parallel Distributed Processing (Hinton 1981, Shastri 1988, Hinton et al. 1986) has suggested that models such as those based on semantic networks could fruitfully be implemented as connectionist networks. In this way, they would retain their useful characterisation of intensional relations but be able to use the powerful content based search abilities of PDP networks (see Chapter 5). Anderson's production system component of ACT would also benefit from the fast search abilities of distributed representations in connectionist networks. Although this work does not specifically deal with attribute binding, it indicates that recent research is searching for new representational media. The work in this thesis develops some tools that allow the characteristics of an actual psychological representational mechanism to be uncovered.

Bartlett's notion of schema has been given a more concrete computational formulation in the work of Minsky, Schank and Abelson on 'frames', 'scripts' and 'dynamic memory' (see Minsky 1977; Schank and Abelson 1977; Schank 1982). A frame or script is a data structure standing for a stereotypical situation. The structure contains slots that can be 'filled in' when the frame is used to remember or understand an actual instance of the situation. The slots may have default assignments which will be maintained if they are not contradicted by the current situation being experienced. Such work has been an interesting attempt to show

how past knowledge can be applied to cognitive processing but, as in the research tradition based on semantic networks, has eschewed consideration of attribute binding. Attribute binding is seen to be the mere filling of a slot with a filler. There is no consideration of how attribute binding might be supported by content. It seems unlikely that Stenning's data can be accounted for by such a simple slot-and-filler device.

Kintsch and van Dijk (1978) presented an influential theory of text comprehension. They claim that text can be represented as a set of separate propositions. These can be written in simple predicate-argument notation. Thus the sentence:

> A series of violent, bloody encounters between police and Black Panther Party members punctuated the early summer days of 1969.

can be represented thus:

- (SERIES, ENCOUNTER)
- (VIOLENT, ENCOUNTER)
- (BLOODY ENCOUNTER)
- (BETWEEN, ENCOUNTER, POLICE, BLACK PANTHER)
- (TIME: IN, ENCOUNTER, SUMMER)
- (EARLY, SUMMER)
- (TIME: IN, SUMMER, 1969)

What Kintsch and van Dijk call 'microstructure' is recovered from the referential relations between propositions. A processor with a buffer of limited size performs one of three basic operation at each time step. It can either link together co-referential propositions or look for a suitable proposition in long-term memory or, as a last resort, make an inference to link a proposition into the microstructure of the text representation. They also discuss the 'macrostructure' of a text — the theme or topic around which the text is organised. As in any purely propositional

model, attribute binding is taken as a representational primitive. Kintsch and van Dijk ignored the attribute binding problem because they were more interested in their conception of text as co-referential propositions made coherent by a macrostructure.

The idea of a search through background knowledge that becomes increasingly more difficult might account for Stenning's semantic ordinal effect but because Kintsch and van Dijk's model used primitives where attribute binding has already been achieved it cannot account for Stenning's data.

Kintsch and van Dijk probably pushed a propositional model as far as it could go — but as Johnson-Laird (1983 pp380-1) points out: it is easy to show that propositions with overlapping arguments need not be co-referential and that propositions with no surface overlap of arguments can be co-referential. Johnson-Laird argues that discourse requires two levels of representation, one propositional and one as a so-called *mental model.* However, mental models do not provide a satisfactory treatment of attribute binding either.

1.4.2 Mental models

Johnson-Laird et al. (1984) criticised semantic network models because they do not deal with extensional relations between entities in the network and the things they represent in the world. Johnson-Laird's work on a representational format he calls 'mental models' (Johnson-Laird 1983) provides an alternative. He claims that as well as some sort of propositional representation, such as a semantic network, people construct mental models of states of affairs in the world. He defines a mental model as a finite and computable representation of information in the outside world as related sets of tokens (Johnson-Laird 1983). An example of the structure of these representations is given by a mental model of a syllogistic premise: All of the X are Y:

$$
\begin{array}{ccc}
x & = & y \\
x & = & y \\
 & & (y) \\
 & & (y)
\end{array}
$$

The model is a representative 'tableau'. The number of tokens corresponding to x's and y's is arbitrary and the items in parentheses represent the possible existence of y's that aren't x's. Johnson-Laird shows how models of premises like these can be integrated to form a model or models that could form the basis of discovering the solution to a syllogism (see Johnson-Laird and Bara, 1984). What is of interest here is the unanalysable structure of the representation of "x is y": x = y. Again we see the that attribute binding is assumed to be primitive and so is not considered to be an important problem with a complex underlying mechanism. It is difficult to see how the construction of a mental model could account for Stenning's data.

1.4.3 Fragmentation Theory

G. V. Jones' 'fragmentation theory' of memory (e.g. Jones 1976, 1978, 1984) is probably the closest model of representational structure in the literature to that described here (see Chapter 3). The theory is based on data from experiments where subjects were given pictures or descriptions of objects (O) which have a particular colour (C) and location (L). Sometimes subjects were also required to recall the sequential position (S) of a stimulus within the presentation set. The errors that subjects make for cued recall are modelled by supposing that memory is based on a fragment or fragments of the attributes in the original description (e.g. a CL and an OS fragment). The status of the links that bind the attributes within a fragment is not discussed.

This paradigm presents only single individuals at a time. When subjects are tested after all the stimuli have been presented they do not have the problem of remembering which attribute belonged to which object within a single presentation. They do have to contend with possible interference between the memories

for different individuals but this is minimised by having non-overlapping descriptions. The experimental paradigm described in this thesis is specifically designed to pose a more demanding binding problem since it uses descriptions of pairs of individuals that overlap considerably both within and across trials.

Fragmentation theory cannot directly account for Stenning's reading time data because it doesn't consider the construction of a representation. If Jones' experimental paradigm were extended to deal with pairs of individuals it might be possible to model the semantic ordinal effect as the increasing difficulty of constructing fragments as more becomes known of an individual. Jones takes the issue of property attribution seriously and develops a parsimonious mathematical model of representation that is in some sense distributed. The model we develop in this thesis is distributed but redundant. Our model is also more complex because the MIT allows us to study more complex stimuli. The comparison between the two models is discussed more fully in Chapter 3.

1.5 The Memory for Individuals Task

The experimental paradigm described in Stenning (1986) has proved to be very flexible and the semantic ordinal effect has generalised over many different experimental manipulations. This section will describe the MIT as it is used in the work described in the rest of the thesis. Some of the variety of experiments that have been performed within the paradigm will then be briefly described.

Most of the experiments performed by our research group including the ones described in later chapters use determinate texts of two individuals classified on four different dimensions. An aspect of the informational structure of the descriptions that was not examined in Stenning (1986) is what we have come to call *matchtype*. The matchtype of a text defines the dimensions on which both individuals have the same or *matching* values and those on which the individuals have contrasting or *mismatching* values. One of the adjective dimensions, usually the most nominal, always mismatches and serves to ensure that two non-identical individuals

Matchtype	Description			
1. + + +	tall	fat	Polish	Bishop
	tall	fat	Polish	Dentist
2. + + −	tall	fat	Polish	Bishop
	tall	fat	Swiss	Dentist
3. + − +	tall	fat	Polish	Bishop
	tall	thin	Polish	Dentist
4. + − −	tall	fat	Polish	Bishop
	tall	thin	Swiss	Dentist
5. − + +	tall	fat	Polish	Bishop
	short	fat	Polish	Dentist
6. − + −	tall	fat	Polish	Bishop
	short	fat	Swiss	Dentist
7. − − +	tall	fat	Polish	Bishop
	short	thin	Polish	Dentist
8. − − −	tall	fat	Polish	Bishop
	short	thin	Swiss	Dentist

Table 1.2: Examples of the Eight Matchtypes. The introducing dimension (e.g. profession) is always mismatched. The other three dimensions can either match(+) or mismatch(−).

are always described. This dimension is known as the *introducer* since it is usually the dimension that is given first in the text. The introducing dimension is usually either an abstract shape such as a square or circle or a profession such as a bishop or dentist. The other three dimensions can either match or mismatch. There are thus eight possible patterns of matching or mismatching and it is these that are called *matchtypes*. Table 1.2 displays the matchtypes for four example dimensions. Matchtypes are referred to by their pattern of matching and mismatching on the non-introducing dimensions, a + denoting a matching dimension and a −, a mismatching dimension. Alternatively, a matchtype is referred to by a number between 1 and 8. The dimensions are named A to D, A corresponding to the introducing dimension. For example, in the vocabulary sets describing people, the dimensions roughly correspond to stature (D), temperament (C), nationality (B) and profession (A).

The vocabulary of a description is usually presented in the reverse of natural order since the first dimension is the nominal. This produces the most natural sounding

description: e.g. There is a bishop. The bishop is Polish. The bishop is fat. The bishop is tall — There is a tall fat Polish bishop.

For convenience, each possible combination of vocabulary items for a given set of four dimensions was given a unique *model number*. Since each dimension is binary and there are eight vocabulary items in a description, the model numbers were all possible 8 bit numbers. In decimal notation there are thus 256 model numbers. If pairs of descriptions that differ only by the order of mention of the individuals are counted as identical, there are 136 possible descriptions for a given set of four vocabulary dimensions. Since there are a large number of possible different combinations of vocabulary dimensions, the total possible number of different descriptions is very large. For a typical vocabulary set of 48 words with binary dimensions (e.g. see Appendix A) the number of possible pairs of individuals is 176,256.

1.6 The Aims of the Thesis

Attribute binding is easy to implement on a serial computer and easy to express in logical notation. Because of this, it has been neglected as a phenomenon worthy of study and experimental paradigms have not been designed so as to pose a *binding problem* to the subject. As Stenning (1986) and much of the work presented in this thesis shows, the problem of attribute binding does, in fact, impose a considerable cognitive load.

Much of the work summarised in Section 1.4 has been concerned with the influence of background knowledge on the representation of new knowledge. Thus, spreading activation ensures that nodes semantically related to those activated are also partially activated and default, stereotypical information in a frame or script is used to "fill in" the missing information necessary to understand the current situation. Work based on semantic networks or frames is essentially limited in the extent to which it can model the influence of content and context by the difficulty that formal logics have dealing with these sources of information (see Stenning

and Oaksford 1989). Because attribute binding has been ignored, the question of how background knowledge might be used to achieve it has not been considered. Since attribute binding is so fundamental, it might be expected that knowledge gained from the structures and processes involved in solving the binding problem might be very useful in the explanation of other aspects of human knowledge representation. Since it seems clear that content is important for cognitive processes, and attribute binding provides a phenomenon intimately dependant on content and yet accessible to study, attribute binding is a useful area to study. It is, of course, an important problem in its own right as a fundamental aspect of knowledge representation. Some of the areas of research for which our model of attribute binding has proved useful can be found in Chapter 6.

The main aim of the work described in this thesis was to develop the necessary methodology for the study of attribute binding. On the way to achieving this aim, a considerable degree of theoretical insight was gained. The models that are developed in the thesis make the claim that the search for associations involving background knowledge is necessary for attribute binding. They also suggest that the representation used is redundant and distributed as well as contentful.

It is clear that attribute binding is a paradigmatic case of human knowledge representation that despite its simplicity requires the use of background knowledge. The work described here shows that it is a tractable problem for experimental investigation and modelling.

1.7 The Structure of the Thesis

This section summarises the content of each chapter in the thesis. The contribution by the author in any team-work described is indicated. There is no isolated 'literature survey' chapter. Instead, background literature is discussed in context in several different chapters.

Chapter 2 discusses the way in which the constructive processes for the repre-

sentation of the pair of individuals in the MIT were probed by measuring the self-paced reading times for each sentence. The chapter discusses the use of this measure and the way in which the data can be modelled using multiple regression techniques. Successful statistical models are described for two experiments. The first 'antonymy experiment' manipulates the degree of antonymy of the vocabulary dimensions and compares free recall to menu recall. This experiment was designed, run and analysed by the author. The second 'replication experiment' collected a large amount of data suitable for recall modelling and showed that the semantic ordinal effect is not accounted for by articulatory rehearsal. The experiment was designed and run by Keith Stenning and Martin Shepherd. The data was analysed and modelled by a team which included the author. The inspiration for the regression modelling sprang from an observation made by the author of the effect of matchtype in the data from the antonymy experiment.

Chapter 3 describes the recall error data from the antonymy and replication experiments. The characteristic error patterns are described and the way in which they were modelled by multiple regression techniques is discussed. The development of the multiple regression techniques was a team effort by the author, Keith Stenning and Martin Shepherd. The author wrote the software package that allowed recall errors to be flexibly classified and different hypothetical features to be defined. The particular regression model described in the chapter was prepared by the author.

Chapter 4 gives a brief description of what is meant by Parallel Distributed Processing (PDP). Some background material is discussed so that the choice of the PDP framework as a whole and the particular network architecture used in Chapter 5 can be motivated.

Chapter 5 describes the PDP networks that were used to model a mechanism by which the feature values of the statistical model could be synthesised into a well-formed recall. The networks were used to model the generation of errors when such representations are disrupted. Such disruption is likely to cause inconsistency which must be resolved if a well-formed recall, even if it is an error, is to be

made. The general PDP modelling approach was developed by the author in collaboration with Keith Stenning. The software package used to simulate the back-propagation networks and the particular models described in the chapter were constructed by the author. All the simulation runs reported were performed by the author.

Chapter 6 gives some concluding remarks and discusses further work that might be fruitful.

The following papers and reports describe various stages in the progress of the above work (Stenning, Shepherd and Levy 1987, Stenning, Shepherd and Levy 1988, Stenning and Levy 1988, Levy and Stenning 1988). The work described here is, at the time of writing, the most fully developed version of the above models.

The MIT, simple as it may seem, has formed the basis of a wide variety of experiments. Apart from those discussed here, other experiments have included one that compared the reading times and memory errors for sets with those for individuals (see Gemmell 1988), one that compared pictorial input to textual input (see Werner 1985), several that examined the use of different text modes and the role that articulatory rehearsal plays (Stenning, Patel and Levy 1987; Patel forthcoming) and several others. Much of the data has still to be fully analysed.

1.8 Theoretical and Methodological Biases

This chapter will be concluded with a summary of some of the theoretical and methodological biases that underly the work reported in the thesis.

We consider that the mechanisms responsible for the representation of which property belongs to which individual are fundamental and non-trivial. Both the processes responsible for the construction of such representations and the underlying structure of the representations are worthy of investigation, theorising and modelling.

Any such representation will be intimately integrated with background knowledge. Any model developed must take this into account. Having said this, it is very hard to model what knowledge a subject or group of subjects have. [2] The work described here attempts to meet the constraints of providing meaningful descriptions that have a well-defined informational structure. The reading times measured in the MIT experimental paradigm give us a measure of the cognitive load affecting the constructive processes at work. As will be seen later in the thesis, both the reading time patterns and the patterns of recall errors can be modelled because they are sensitive to the informational structure of the descriptions given. The theoretical interpretation given to the statistical models must include an explanation for the use of the knowledge-base.

The materials used in the MIT represent a compromise between the opposing approaches of Ebbinghaus (1885, 1965) and Bartlett (1932). Ebbinghaus used materials devoid of content to allow objective measurement. His methodology has been largely rejected in recent times because his tasks are not representative of everyday cognition and it can be shown that subjects will impose their own unpredictable and individual associations on nonsense material to make it memorable. The Bartlettian tradition stresses the necessity for natural materials and tasks. The drawback is the difficulty of consistent measurements and explicit models that have explanatory adequacy (Chomsky 1965, Johnson-Laird 1983). The tension between the two approaches is summarised by Baddeley (1976, p14):

> The study of memory continues to be torn between Ebbinghaus's insistence on simplification (with its attendant dangers of trivialization) and Bartlett's emphasis on the complexities of human memory (with its danger of intractability).

The descriptions used by the MIT are meaningful and allow subjects to import their own background knowledge in an *effort after meaning.* However, the regularities in the structure of the descriptions allow measurement and modelling.

The MIT is unusual in allowing two relatively independent measures to be taken

[2] An attempt within the MIT was made by Nelson (1988)

for the same representational process. It is to be hoped that the model of reading time data and the model of recall error data will combine synergistically to give a better understanding of the representation of individuals than either model alone.

Much of the work reported here takes an exploratory model building approach rather than strict Popperian hypothesis testing. This is not to say that the experiments are not designed with hypotheses to test but rather that the data we obtain is so rich that it demands to be modelled. It is usually the case that the models discussed in this thesis have been strengthened by their applicability to the data from further experiments done within our research group.

We find that Parallel Distributed Processing (PDP) is a useful modelling framework. Although not yet used in the context of the reading time data, it proved to be a fruitful method of extending the statistical model of recall that we constructed. As will be discussed in Chapter 4, the PDP framework is a fruitful source for metaphors and concrete models for cognitive processes and representations.

Chapter 2

Modelling Reading Times

2.1 Introduction

The self-paced sentence-by-sentence reading component of the MIT is designed to collect a separate reading time for each of the eight sentences in a description. This chapter describes the statistical methods used to describe how the reading times of each sentence varied. This statistical exercise is theoretically interesting because we claim that reading times reflect the cognitive loads imposed by the incremental interpretation of the descriptions. This chapter describes how the results of two experiments replicate and extend the initial result of Stenning (1986) that reading times for sentences concerning an individual tend to increase as the individual becomes increasingly specified. Observations of the reading time curves for the data from the first experiment inspired the use of a multiple regression modelling framework which was successfully used by the whole research group. The use of this statistical method for the modelling of a large amount of data from a further experiment is then described. The chapter finishes with a note of the implications of the general reading time model for a theory of how the binding problem is solved and individuals are represented within the confines of the MIT.

A full discussion of the recall errors from both experiments and their analysis and modelling is found in Chapter 3.

	Reading Time of target sentence (sec)
Condition 1 — No role change:	
John was not looking forward to teaching maths.	
The bus trundled slowly along the road.	
He hoped he could control the class today.	1.72
Condition 1 — Role change on target:	
John was on his way to school.	
The bus trundled slowly along the road.	
He hoped he could control the class today.	1.89

Table 2.1: Materials and target reading-time results from Sanford and Garrods' (1981) role-shift study

The use of reading time as a measure of cognitive load is a widespread methodological tool. Some representative examples in the literature that have some bearing on the work reported here are the work of Sanford and Garrod, and Kieras.

Tony Sanford and Simon Garrod have performed many interesting experiments that involve the measurement of reading times (see Sanford and Garrod 1981, Sanford 1985). Reading time results are used to infer how various knowledge sources are used in the comprehension of written text. An example is an experiment reported in Sanford and Garrod (1981) where the reading times of two target sentences demonstrate that if the perceived role of an individual in a text changes the text becomes harder to understand. An example pair of texts is shown in Table 2.1. Sanford and Garrod's interpretation of this result is that a crucial aspect of the processing of the text is recruitment of general knowledge pertaining to the role of the topic individual. If a subject is fooled into assigning the wrong role she experiences some difficulty when this mistake is discovered because new structures in memory must be tapped to interpret the text. This difficulty is revealed by an increase in reading time.

Sanford and Garrod's concerns contrast rather sharply with the work reported here. The work in this thesis *is* concerned with recruitment of general knowledge needed to understand a description. However, we see the need to first establish

the fundamental aspects of representation concerned with attribute binding. The reading times reported here do reflect the use of background knowledge but at this point in the research we are not concerned with distinctions between knowledge sources. What is seen to be important in the experiments described here are structural aspects of the description that control the difficulty of the representational task, by affecting how much background knowledge needs to be called upon. Probing the detailed representation of specific knowledge is a hard problem. A start in characterising subjects' general knowledge of the descriptions used in this experimental paradigm has been made by Nelson (1988).

Kieras (1981, 1984) describes experiments where self-paced reading times were measured for the sentences of simple passages. The coherence of the passages was manipulated by changing the ordering of the sentences. High coherence passages would contain more *given* referents than *new* referents (Haviland and Clark 1974), allowing easier integration of the information of new sentences into the representation of the previous sentences. Passages also varied as to whether or not they began with a good 'topic sentence'. Multiple regression was used to assess the relative contributions to reading times of various hypothesised aspects of representational processing. The methodology of Kieras' work has some strong similarities to the work described in this chapter. However, the material used is more complex and the work aims at accounting for the integration of complex propositions rather than the simpler and more fundamental problems of attribute binding.

Reading times are an attractive psychological measure because they have proven to be sensitive to such a wide variety of different factors. Multiple linear regression is a useful statistical technique for the modelling of reading times because it allows the influences of a large number of predictor variables to be easily quantified. A relevant selection of articles can be found in Kieras and Just (1984). In that book Knight (1984), Haberlandt (1984), and Graesser and Riha (1984) review some of the issues involved in the use of multiple regression techniques and their application to modelling reading time data.

2.2 The 'Antonymy Experiment'

This section describes an experiment designed to generalise the applicability of the MIT. Rather than collecting large amounts of data to allow full scale modelling of reading times and recall errors (see Section 2.3), it was designed to find out whether the general findings of the MIT would still hold if some of the rather rigid restrictions of this experimental paradigm were relaxed.

Most experiments carried out within the MIT paradigm have used binary property dimensions. These contained properties that were always paired, so that they always appeared together in a recall menu and were often antonymous (e.g. fat vs. thin, or mad vs. sane). The experiment compared this type of description to one in which the property dimensions were less arbitrarily 'binary'.

The other main experimental manipulation was carried out to check that the data from the MIT did not depend on using a menu recall task. Menu recall was compared to a free recall task. Both tasks were performed after a variable period of counting backwards in threes, designed to ensure that the recall tested was fairly 'deep' or 'semantic' and not merely testing a surface representation. If this were so, it was reasoned that recall should be fairly good, even after a delay, and the variable delay should have little differential effect on recall performance (see Chapter 3).

2.2.1 Experimental Details

Subjects

There were 12 paid student subjects.

Materials

The texts were all generated in Franz LISP on a VAX minicomputer. They were all descriptions of two individuals classified on four adjective dimensions using a Property by Property mode. The vocabulary dimensions roughly corresponded to occupation, nationality, physical character and temperament.

(a) 'Binary' Texts:- The texts were generated by randomly picking one pair per cohort from a fixed list of binary pairs. (See Appendix A for the actual materials). This is the method used to generate the vocabulary for most of the experiments performed within the MIT paradigm. The word pairs constituted the 'setting' of the text and went to make up the recall menu if such a menu was used. e.g:

(dentist/baker) (Swedish/Russian) (hungry/thirsty) (sane/mad)
There is a dentist
There is a baker
The dentist is Swedish
The baker is Russian
The dentist is hungry
The baker is hungry
The dentist is sane
The baker is sane
There was a sane hungry Swedish dentist and a sane hungry Russian baker

This example is the first full text that was used in the experiment. The first line is the setting, then come the 8 sentences of the PxP text and the final line is the feedback sentence presented to the subject after the recall stage. The matchtype here is $++-$ (see Table 1.2).

In the binary texts the pairs were fixed so that, for example, the nationalities 'Swedish' and 'Russian' were always associated. Some of these pairs were opposed e.g. (mad/sane) but others, such as the nationalities, were just non-overlapping pairs, i.e. although the two adjectives in a dimension were not opposed, one indi-

vidual could never be described with both adjectives. [1]

(b) 'Non-binary' Texts: These texts were generated for each cohort by randomly choosing one word from each of two lists (See Appendix A). The lists were constructed so that the pairs used for the binary texts occurred within the lists rather than between them. Thus none of the oppositions that occurred in the binary texts could occur in the non-binary ones. The vocabulary set as a whole was the same as that for the binary texts. The pairs for each cohort of each text were generated randomly for each text so that pairs were not associated, although the small size of the vocabulary set meant that pairings could quite often be repeated. The texts had no settings so that there was not even this slight prior basis for associating the members of a vocabulary dimension.

e.g. :

(vicar/teacher) (German/Russian) (young/hungry) (sane/weak)
There is a vicar
There is a teacher
The vicar is German
The teacher is German
The vicar is young
The teacher is hungry
The vicar is sane
The teacher is sane
There was a sane young German vicar and a sane hungry German teacher

In this example the first (setting) line serves to show the cohort pairs but was not in fact presented to the subject.

[1] Unusual cases such as dual nationality were ignored.

Reading Task

The Reading, Counting and menu Recall tasks were all performed on a BBC ECONET microcomputer network.

Before each session the subject was given printed instruction sheets to read. There were four different instruction sheets each appropriate to the four combinations of reading and recall tasks. After the sheets were read the subject was given five example texts to read and recall. The practice texts used an object-based vocabulary rather than the people-based one used in the experiment. The subjects were told to perform as quickly as possible while being as accurate as possible.

Counting task

Between reading and recalling the text subjects were required to count backwards aloud in threes from a four-figure number (randomly chosen from between 2000 and 7000). This task could last five, ten or twenty seconds and subjects were instructed to count in time with a tone that occurred every second. See Brown (1958), Peterson and Peterson (1959) and Baddeley (1976) for the origins of this sort of rehearsal disrupting task.

Recall task

(a) Menu recall:- After the Counting task subjects were asked to recall the people that had been described in the text by picking them out from a menu. The menu was constructed from the setting (whether it had been presented or not) and words were selected by moving a cursor. There were no restrictions on the order of recall of people or of the words describing one person, but a separate menu was used for each person. Editing of the recalled person was possible. An example menu looked like this:

Recall one of the people

sane	weak
young	hungry
German	Russian
vicar	teacher

After both people had been recalled the subject was shown a feedback sentence which was a correct description of the people presented in the text. This was presented for 4 seconds.

(b) Free recall: After the counting task the subject was asked to recall aloud the people described by the text. Any unambiguous description was allowed. The recall was noted down verbatim by the experimenter. A feedback sentence was then shown.

Design

There were two groups of six subjects. One group received binary texts and the other non-binary ones.

Each subject performed four sessions - two free recall and two menu recall. These sessions were alternated and the order of recall blocks balanced within each group, so that three subjects did a 'menu-free-menu-free' ordering and the other three did a 'free-menu-free-menu' ordering. Each session contained 24 randomly ordered texts — the 8 matchtypes crossed with the three recall delays. Each session was divided arbitrarily into two subsessions of 12 texts, allowing a short rest between subsessions if required.

Thus the binary/non-binary factor was between subjects. For each group the recall, matchtype and delay factors were crossed within subjects. So within each group the design was 2 x 8 x 3 = 48, and each subject completed 2 whole designs

		Individual 1				Individual 2			
Sentence		1	2	3	4	1	2	3	4
BINARY	menu	1.36	1.94	2.05	2.16	1.38	2.28	2.25	3.52
	free	1.60	2.32	2.41	2.69	1.66	2.75	2.93	4.59
NON-BINARY	menu	2.30	2.28	2.71	3.32	2.78	3.90	4.53	7.08
	free	3.01	2.84	3.70	5.12	3.53	4.79	6.19	9.47

Table 2.2: Mean reading times (sec) as a function of text type and recall task

within her group = 96 texts.

The sessions lasted between 30 and 60 minutes.

2.2.2 Descriptive Results

This section serves to give a general description of the reading time curves as well as the descriptive statistics based on the experimental design.

An ANOVA was performed using the following design specification:

binary/non-binary	(2 levels)	- group factor i.e between subjects
subjects	(6 levels)	- random factor
Matchtype	(8 levels)	
free/menu recall	(2 levels)	
individual	(2 levels)	
sentence	(4 levels)	

The data was trimmed to reduce the effect of extreme reading times, the procedure being as follows: the standard deviation was calculated for each subject and for each sentence. Scores above or below 2 standards deviations from the mean were trimmed to this cutoff point. Nearly all of the 4.6% of the data that was trimmed was above the upper cutoff point rather than below the lower one.

Results of ANOVA

- There was a main effect of sentence ($p = 0.0002$):

sentence	1	2	3	4
mean reading time	2.20	2.89	3.35	4.74

This convincingly replicates the Stenning (1986) Semantic Ordinal Effect.

- The main effect for individual was just insignificant ($p = 0.0553$): individual 1 = 2.61; individual 2 = 3.98
- There was an interaction between individual and sentence ($p = 0.0218$).

  These two results indicate a probable difference in the way the two individuals are processed.

- The main effect for group (bin/nonbin) was insignificant ($p = 0.0659$): bin = 2.37; nonbin = 4.22

  This statistical result was surprising and probably brought about by having too few subjects since this was a between subjects variable and reading times are notoriously noisy data. Despite the near miss for significance, the binary and non-binary data was analysed separately in the regression analyses below to see whether different loading patterns emerged. This was felt to be justified by the large numbers of significant interactions involving the binary/non-binary variable.

- There was a significant main effect of recall task ($p = 0.0175$): menu = 2.86; free = 3.73

  It seems that subjects found that recalling without a menu required that the texts were read with more effort. This is interpretable since free recall requires that the vocabulary items themselves be remembered whilst the menu provides the vocabulary items as cues for the assignment of adjectives to individuals (attribute binding).

- There was an interaction between recall type and sentence ($p = 0.0127$).

  This result is harder to interpret. There is no obvious reason why the extra load imposed by free recall should have a different effect for the different sentences making up the description of an individual. It appears that the mechanism responsible for the semantic ordinal effect interacts with the effect of different recall tasks.

- There was a main effect of matchtype ($p < 0.0001$):

matchtype	1	2	3	4	5	6	7	8
mean reading time	2.66	3.03	3.14	3.38	3.20	3.53	3.63	3.79

Matchtype is clearly an important determiner of reading times. There were also several interactions involving matchtype:

- There was an interaction between group (binary/non-binary) and matchtype ($p = 0.0095$).
- There was an interaction between matchtype and individual ($p = 0.0074$).
- There was a interaction between matchtype and sentence ($p < 0.0001$).
- There was an interaction between group (binary/non-binary), matchtype and sentence($p = 0.0110$).
- There was a interaction between matchtype, individual and sentence ($p < 0.0001$).
- There was an interaction between group (binary/non-binary), matchtype, individual and sentence ($p = 0.0022$).

These six results demonstrate the complex effect of matchtype on reading times and suggest that it is an important factor in the mechanism underlying the semantic ordinal effect. It also appears that the binary/non-binary manipulation is more important than its non-significant main effect would suggest.

It is not clear from these descriptive results why matchtype is so important. The observation that the reading time curves for the different matchtypes had radically different shapes (see Figure 2.1) inspired the use of multiple regression as a technique to model rather than simply describe the data.

2.2.3 Multiple regression modelling

The regression model was built on a number of observations and assumptions. The semantic ordinal effect is an observation that reading times increase for a sentence

describing a particular individual as the number of previous sentences about that individual increases. The reading times seem to be affected by the matchtype of the description — the reading time curves have consistently different shapes for the different matchtypes (see Figure 2.1). In particular, it seems that a sentence which establishes that a particular dimension mismatches takes longer to read than one which establishes a match. This makes sense in terms of the different amounts of information imparted by matched and mismatched dimensions. A matched dimension might be remembered as a single lexical item; a representation of the word 'fat' would be enough to represent the fact that both the dentist and the bishop were fat. If, however, one of them was fat and the other was thin, it would be necessary to represent *which* one was fat and *which* one was thin, as well as remembering that they differed along this dimension.

We have assumed that the semantic ordinal effect is due to an increasing cognitive load that is manifested by an increased reading time. The cognitive load is mostly imposed by constructive processes rather than maintenance processes such as articulatory rehearsal [2]. The constructive processes build a representation of the attribution of properties to individuals. The ANOVA has shown that a major determiner of reading time (our measure of cognitive load), is simply the match structure of the descriptions. What remains is to break down the cognitive load into different component processes by modelling the way in which different aspects of the matching structure affect reading time.

It is clear from the results of the ANOVA and the shape of the reading time curves that reading times tend to increase as more is known about an individual and that matchtype affects the shape of this increase. It is also clear that sentences that establish a mismatch take longer to read than those that establish a match. From these observations it is reasonable to hypothesise that reading time is dependant on the accumulation of cognitive load from previous matching and mismatching dimensions as well as local loads resulting from establishing the character of the present dimension. These hypotheses give rise to variables that track the number of previous matches and mismatches as well as binary variables that describe

[2]This is demonstrated in Section 2.3.2

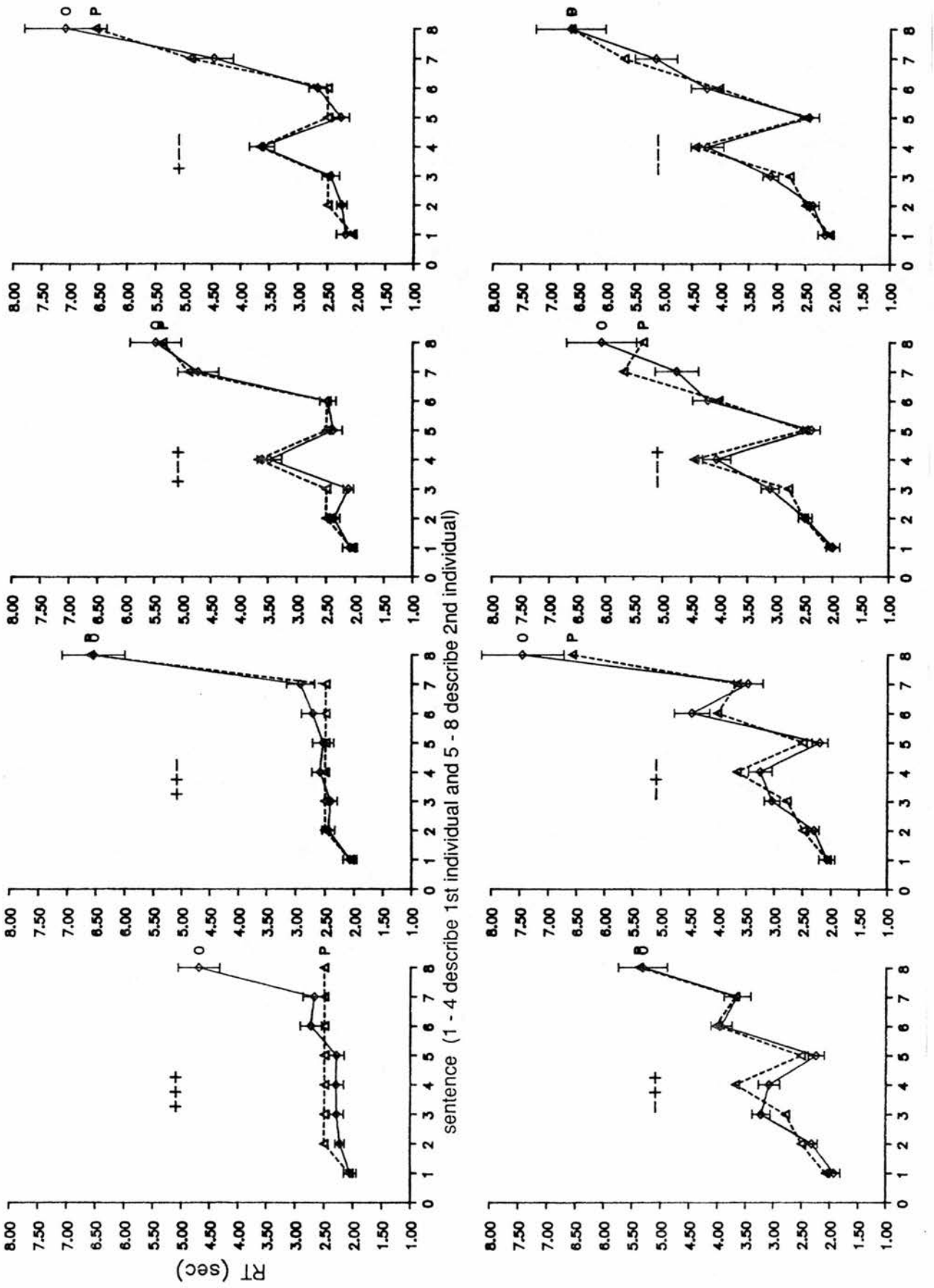


Figure 2.1: Observed and Predicted Reading Times for the Antonymy Experiment

whether the current sentence establishes a match or mismatch. Lastly, there are likely to be factors that are in some sense precomputed because they can be confidently predicted to be present for all descriptions. The use of multiple linear regression to model the different loading factors imposes the restriction that the processes are independent and additive.

Multiple regression allows a progressive modelling of the data. Using procedures such as 'best possible subsets' regression and 'stepwise' regression (see Draper and Smith 1981, Dixon et al. 1983), the statistical procedure itself can be used to select the variables that provide the best account of the variance in the data, as well as assigning each a coefficient or slope. Different possible variables can be defined and 'tried' out on the data. The residuals (differences between the observed data and 'predicted' model) can be used to define better independent variables. This approach is open to the abuse that variables may be chosen purely to fit the data without theoretical justification. The models reported here are based purely on the structure of the descriptions and the variables are sensibly interpretable. Furthermore, this danger is counteracted in the research described here by showing that the same form of model can account for data from completely different experiments.

Two types of variable were proposed. The first type was an accumulative load; the second type was a local non-recurrent load. The first type consisted of the two variables, MISLOAD, the accumulated number of mismatching dimensions that had been encountered, and MATLOAD, the accumulated number of matching dimensions. These variables are incremented when the sentence describing the second individual along a particular dimension is read, since it is only then that a match or mismatch is discovered. The second type of variable is represented by LOCMIS, a binary variable that takes a value of 1 when a mismatched dimension is discovered. LOCMIS takes into account the processing needed to represent the correct attributions of a mismatched dimension. It is assumed that there is no local load for the mismatching introducing dimension since it invariably mismatches and so LOCMIS is always zero for the first dimension. The first dimension still contributes to MISLOAD, however, since the representation of

Variable	Value				
MISLOAD	0	1	2	3	4
MIS1	0	1	0	0	0
MIS2	0	0	1	0	0
MIS3	0	0	0	1	0
MIS4	0	0	0	0	1

Table 2.3: The definition of MISLOAD in terms of four dummy variables

further dimensions has to be integrated with its representation.

The argument made above that mismatched dimensions are significantly harder to represent than matched dimensions is born out by the fact that LOCMIS is included in the model but 'LOCMAT' is not selected, even if it is available to the regression procedure. This is part of the explanation of why a mismatched dimension takes longer to read than a matched one. This longer reading time is also accounted for by larger coefficients for MISLOAD than for MATLOAD.

A striking aspect of the reading curves is the flatness of the curves for the first two matchtypes (see Figures 2.2 to 2.5). It seems that the matched dimensions here impose very little load. A plausible hypothesis, and one that fitted the data, was that MATLOAD does not take effect until a mismatched dimension (apart from the invariably mismatching introducers) has been met. After such a mismatch it is claimed that MATLOAD takes full retrospective effect. It is assumed that matched dimensions are particularly easy to process before the representation of a mismatch forces processing to focus on separate individuals. This constraint restricts the value of MATLOAD to 0, 1 or 2.

A method of estimating the the shapes of the slope of each variable is to define a separate binary dummy variable for each level of the load variable apart from zero. As an example, Table 2.3 defines the four dummy variables for MISLOAD.

The values of each variable at each point in the text are summarised in Table 2.4. For the sake of completeness, the table contains the definitions for both IxI and PxP texts, although for the moment we are only concerned with the PxP definitions. The table also contains the definition for a variable, NEUTLOAD, that

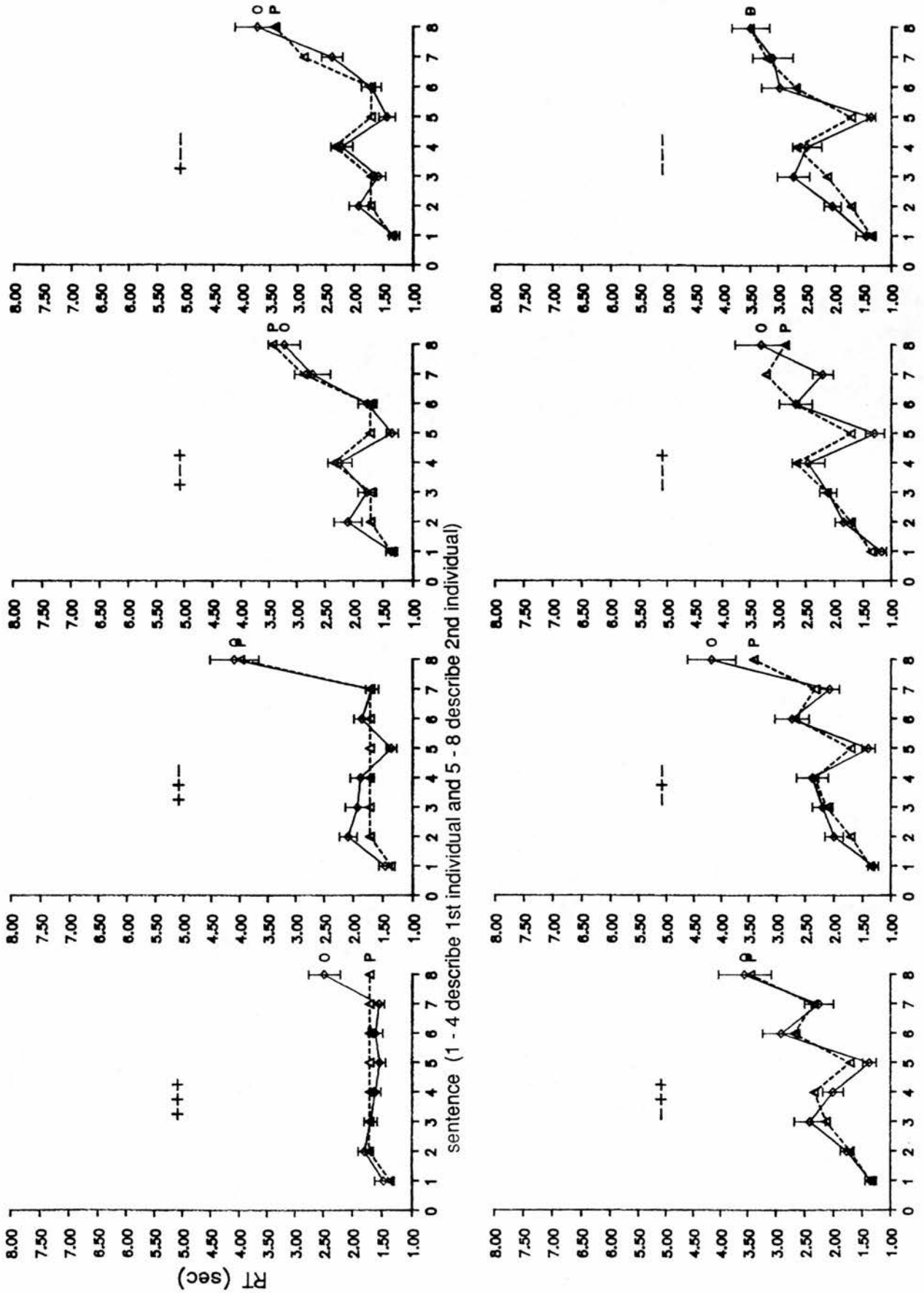


Figure 2.2: Observed and Predicted Reading Times for the Binary Menu Data

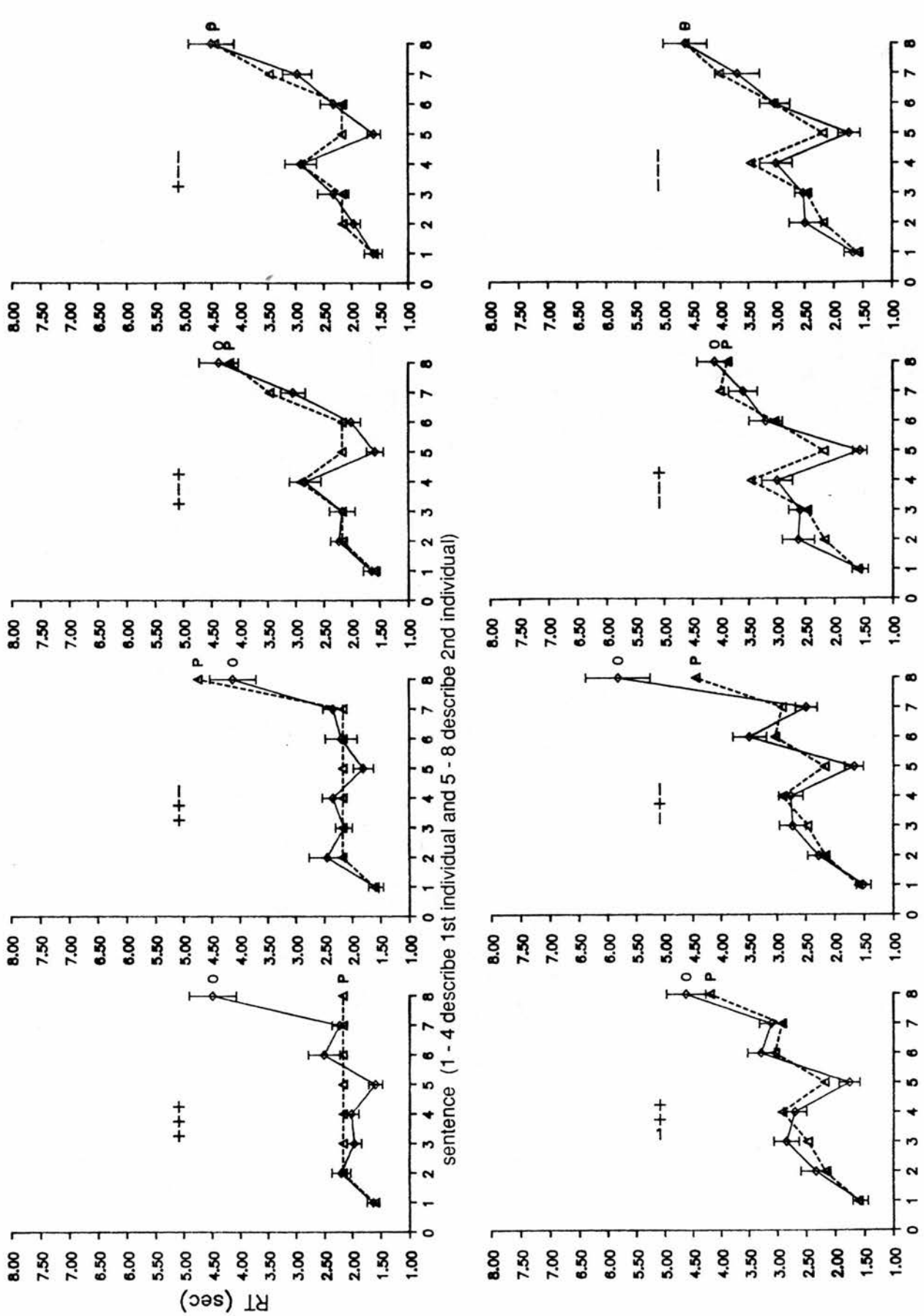


Figure 2.3: Observed and Predicted Reading Times for the Binary Free Data

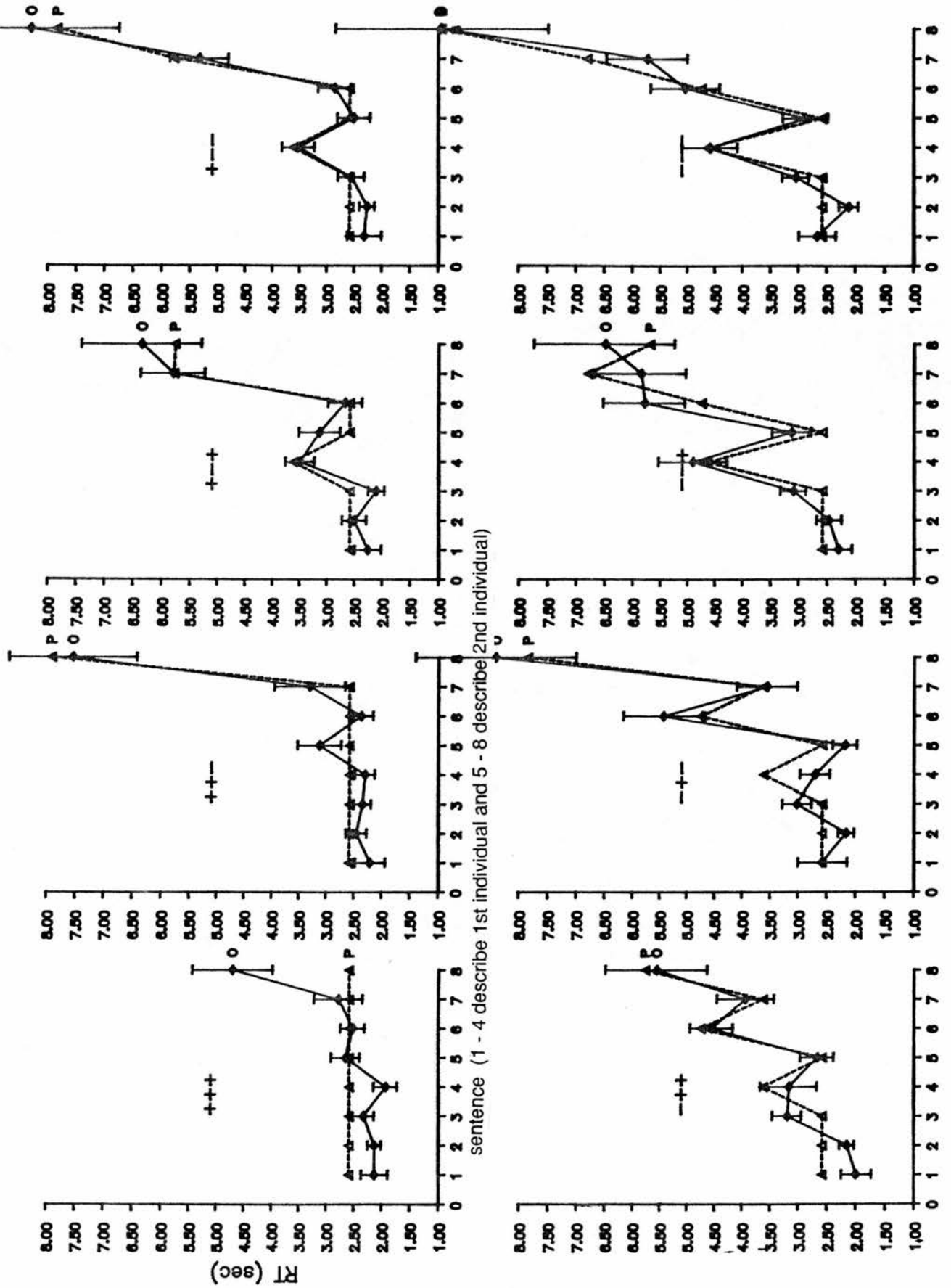


Figure 2.4: Observed and Predicted Reading Times for the Non-binary Menu data

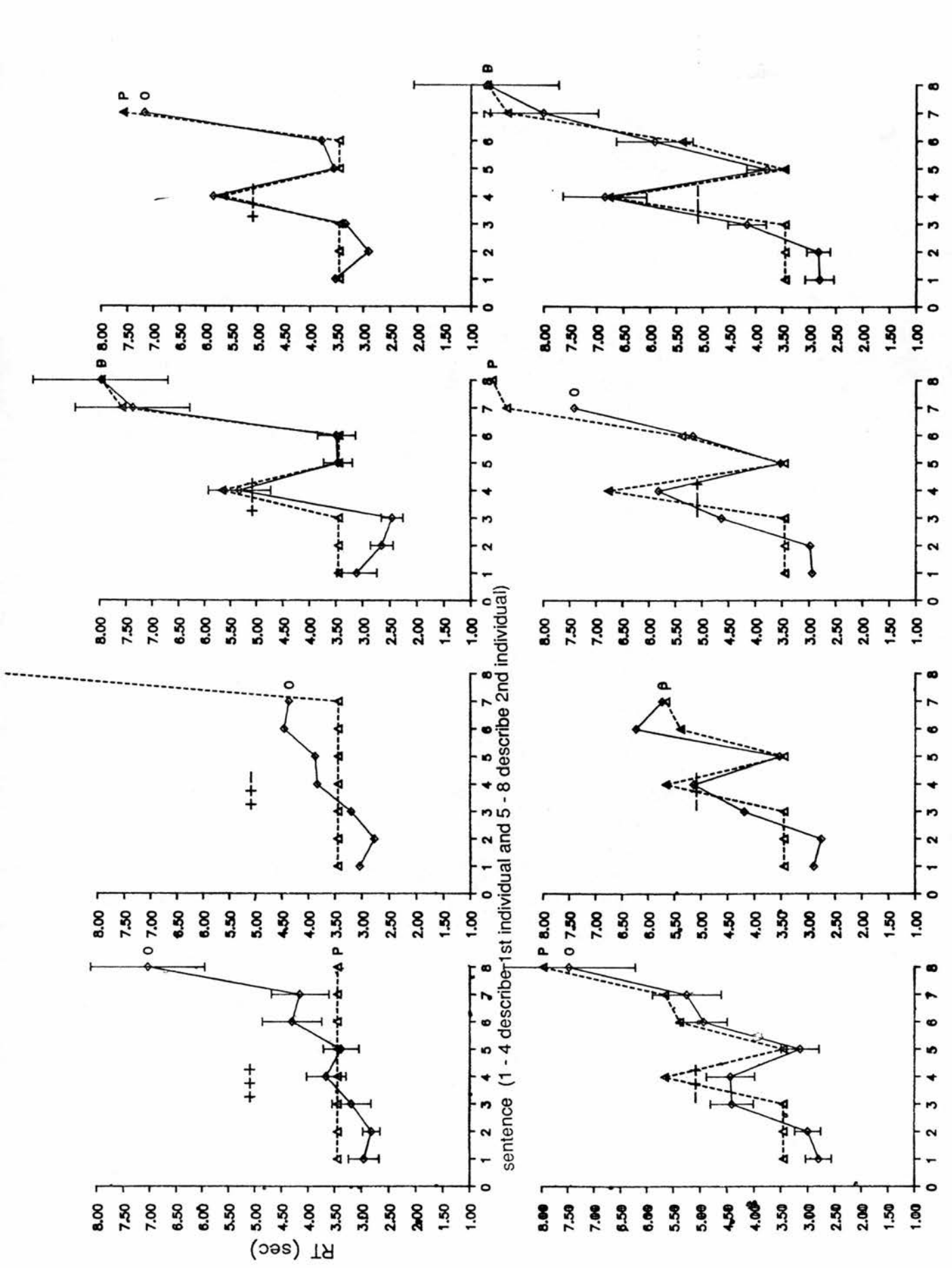


Figure 2.5: Observed and Predicted Reading Times for the Non-binary Free data

	Sentence							
	1	2	3	4	5	6	7	8
	N+-L	N+-L	N+-L	N+-L	N+-L	N+-L	N+-L	N+-L
IxI:								
+ + +	1000	2000	3000	4000	0010	0010	0010	0010
+ + −	1000	2000	3000	4000	0010	0010	0010	0221
+ − +	1000	2000	3000	4000	0010	0010	0121	0220
+ − −	1000	2000	3000	4000	0010	0010	0121	0131
− + +	1000	2000	3000	4000	0010	0021	0120	0220
− + −	1000	2000	3000	4000	0010	0021	0120	0131
− − +	1000	2000	3000	4000	0010	0021	0031	0130
− − −	1000	2000	3000	4000	0010	0021	0031	0041
PxP:								
+ + +	1000	0010	1010	0010	1010	0010	1010	0010
+ + −	1000	0010	1010	0010	1010	0010	1010	0221
+ − +	1000	0010	1010	0010	1010	0121	1120	0220
+ − −	1000	0010	1010	0010	1010	0121	1120	0131
− + +	1000	0010	1010	1020	0120	1120	1120	0220
− + −	1000	0010	1010	1020	0120	1120	1120	0131
− − +	1000	0010	1010	1020	0120	0031	1030	0130
− − −	1000	0010	1010	1020	0120	0031	1030	0041

Key:N, NEUTLOAD; +, MATLOAD; -, MISLOAD; L, LOCMIS

Table 2.4: The Definitions of the Regression Variables for IxI and PxP Text Modes. The table shows the value of the four load variables at each sentence for each matchtype.

will be discussed in Section 2.3.3. I will now describe several regression models for the data described above. The first is a general model for the whole data set. The following four models account for the data split by the levels of the binary/non-binary and menu/free variables of the experimental design. In all cases, the models themselves are statistically significant as are all the individual variables. The statistical procedure used was 'best possible subsets' multiple linear regression using the P9R program of the BMDP statistical package (Dixon et al. 1983).

Variable	Coefficient	Std. Error
Intercept	4.28	0.14
MIS1	0.43	0.10
MIS2	0.72	0.13
MIS3	2.39	0.16
MIS4	3.34	0.29
MAT1	0.89	0.11
MAT2	2.59	0.17
LOCMIS	1.24	0.10
BIN/NON-BIN	-1.85	0.06
MENU/FREE	-0.86	0.06

Table 2.5: The General regression Model

Partition	% Variance
Pure Error	73.0
Regression	22.6
Lack of Fit	4.4
'Success'	84.0

Table 2.6: Variance Partition for General Model

General model

A model of the whole data set serves to show that the basic approach is appropriate. As well as the variables defined above, this model contains binary variables to account for the binary/non-binary and menu/free recall distinctions. The statistical model is summarised in Table 2.5, Table 2.6 and Figure 2.1.

Table 2.6 summarises the method used to judge the success of the regression modelling. Much of the variance in the data is noise or 'pure error' - this is calculated as the variance between 'repeated measures', i.e. the difference between reading times for identical levels of the independent variables (see Draper and Smith 1966, p. 33). It is impossible to account for any of this variance using the variables in the model, so the degree of 'success' is taken to be the proportion of the remaining variance accounted for by the model. Thus $success = R^2/(total\ variance - pure\ error)$, where R is the multiple correlation coefficient of the model and R^2 is the amount

of total variance accounted for by the model. [3] This is merely a method of reporting the results and is not a substitute for obtaining a statistically significant model with significant and interpretable coefficients.

Of course, some of this 'pure error' probably contains some interesting variance that could, in principle, be accounted for in a analysis using other variables not distinguished here. If one could collect enough data, differences between materials might be measurable. However, within this paradigm, it would take a truly enormous experiment to gather sufficient data for this kind of analysis. Differences between subjects are potentially interesting if they reveal differences in cognitive strategies. An idea of the amount of variance in the data due to inter-subject difference can be gained by adding a 12 level variable for the different subjects to the model and noting how much the pure error measure is reduced. If this is done for the above data it appears that 41% of the total variance is caused by inter-subject factors. Individual subject analyses are not practically possible due to sparsity of data but the reading time curves conform roughly to the regression model.

The general model produces a fair account of the data. The fit is impaired by the fact that the differences between the four quarters of the data are not captured by two simple binary variables since, as can be seen by many of the interactions in the ANOVA, the reading time curves differ in shape as well as relative magnitudes.

The load variables, MISLOAD and MATLOAD increase at each level as would be expected if they are estimating an accumulating cognitive load. The coefficients for MAT1 and MAT2 seem surprisingly high compared to MIS1 - MIS4 until it is remembered that many matching dimensions do not contribute at all to MATLOAD at the time of reading. MAT2 is high because it only comes into effect for sentence 8 and is absorbing some of the large reading times that occur here. LOCMIS appears to impose a significant non-recurrent load. The binary/non-binary and menu/free recall variables unsurprisingly support the fact that binary texts are read faster than non-binary ones and texts that are known to be followed

[3] All R^2 statistics quoted are 'adjusted R^2' statistics.

by a menu are easier than those followed by a free recall test.

The largest residual is the one produced for the last sentence of matchtype 1 (see Figure 2.1). This matchtype has a particularly good 'figure' since all its non-introducing dimensions are matched. It is perhaps not surprising that it does not follow the more general assumptions that apply to the other matchtypes.

The general analysis establishes that the modelling method is successful in estimating the slopes of the proposed loading variables. It would be expected that even more successful models would be obtained from the individual quarters of the data since the bin/non-bin and menu/free variables interact with several of the other variables in the ANOVA and clearly have an effect on the shape of the reading curves.

The Four Quarters

As expected, the regression models for the four quarters of the data are more successful in accounting for the variance in the data than the general model. The differences between the four models are easily interpretable. The reading time curves, together with the 'predictions' of the models are displayed in Figures 2.2 to 2.5.

The intercept term in the regression models can be interpreted as reflecting 'overhead' processes common to all sentences. The intercept rises in the order Bin Menu, Bin Free, Nonbin Menu, Nonbin Free. This is evidence that the overhead processes are sensitive to the overall difficulty of the task.

Generally speaking, the coefficients in the models increase in the order Bin Menu, Bin Free, Nonbin Menu, Nonbin Free. The exceptions are MIS4 and LOCMIS for which the coefficient for Nonbin Menu is greater than the corresponding one for the Nonbin Free model. The most obvious differences between the models are the missing MIS1 and MIS2 variables for the non-binary data. This is probably due to the difference between the shallow reading times for the first three sentences

Variable	Bin Menu	Bin Free	Nonbin Menu	Nonbin Free
Intercept	1.36	1.60	2.58	3.45
MIS1	0.36	0.57	–	–
MIS2	0.77	0.88	–	–
MIS3	1.31	1.86	2.06	3.32
MIS4	1.60	2.47	4.62	3.71
MAT1	0.21	0.44	1.04	2.21
MAT2	1.32	1.72	3.18	4.54
LOCMIS	0.55	0.56	2.15	1.94

Table 2.7: The Regression Models for the Four Quarters of the Data

Partition	Bin Menu	Bin Free	Nonbin Menu	Nonbin Free
Pure Error %	79.2	75.1	80.3	81.1
Regression %	19.1	23.0	19.0	18.3
Lack of Fit %	1.7	1.9	0.7	0.6
'Success' %	91.8	92.3	96.4	96.8

Table 2.8: Variance Partition for the Four Quarters

of the first individual [4] and the extremely steep curves for the second individuals. The missing MISLOAD levels coupled with the much higher values for LOCMIS for the non-binary data point to a strategy where although there is a local effect when a mismatch is detected, the recurrent load takes effect towards the end of the text. The strategy that subjects use for non-binary text appears to delay some components of constructive processing until later in the text than that used for the binary material.

An interesting part of the results is that the local processing load represented by the LOCMIS variable is sensitive only to the Binary/Non-binary distinction and not to recall task. It appears that the increase in cognitive load produced by the expectation of a free recall test affects only the recurrent load processes.

The MAT1 and MAT2 variables have much higher coefficients for the non-binary material than for the binary texts. This is further evidence of the greater difficulty subjects have with the non-binary descriptions.

There is a similar failure to model the raised reading time for the final sentence of

[4]There is even some evidence of a *downwards* trend in reading time for the third matchtype of the non-binary free recall data.

matchtype 1 (fully matched texts) in the models for the four quarters as there was in the general model. Evidently there is some effect of MATLOAD here although it can be delayed until the very end of the text. This consistent residual may be evidence that there is some special end-of-text processing done after sentence 8. For most matchtypes, this is absorbed by the coefficients in MAT2 and MIS4, but this can't happen in matchtype 1.

The most interesting result of comparing these four models is the greater difficulty that subjects have with free recall as compared to menu recall and non-binary compared with binary material. The difference between the models for free recall and menu recall is relatively small and seems easily interpretable as a greater effort needed to retain the vocabulary items themselves as well as the mnemonic support for the correct assignments of adjectives to individuals. For a menu recall it suffices to remember, for example, that the bishop was Polish and the dentist wasn't but free recall may also require the subject to remember that the dentist was Swiss. This difficulty is compounded for the non-binary texts since there was no constant relation or association between the pairs of vocabulary items of a dimension.

The difference between the models for binary and non-binary texts is large. It reflects the inherent difficulty that non-binary dimensions present to subjects within this experimental paradigm. It is simply harder to recruit general knowledge for non-binary texts because they have less regularity in the way vocabulary dimensions are generated. It is interesting that a difference in the way the content of the members of the same dimension can be related is so successfully modelled by a method that seems so drily content-free and structural in nature.

Although the binary menu data and binary free data are described by the same combination of variables, this does not mean that the same statistical model suffices to describe them both. An F-test of the equality of regression lines across both subsets of data performed by program P1R of the BMDP package showed that the equations describing the two subsets were significantly different ($F(8, 4592) = 20.325; p < 0.001$). The same was true of the two subsets of

the non-binary data ($F(6, 4596) = 23.426; p < 0.001$). Thus, although there are qualitative similarities between the members of these pairs of models, they differ quantitatively.

Since the non-binary material is perhaps more 'realistic' than the binary texts, if we are to continue using binary material in the MIT, it is important that the form of the regression models for binary and non-binary descriptions is the same. Although the reading times for the non-binary texts are much longer, the same variable definitions are able to account for both the non-binary and the binary data.

2.2.4 The Relationship between ANOVA and multiple regression

Although the ANOVA is a special case of multiple regression and can be extended to give much the same information, the two methods are typically used for very different purposes (see Carpenter 1984). ANOVAs are convenient methods of showing whether or not nominal variables or interactions between them have an effect on a dependent variable. Multiple regression typically assigns each independent variable a quantitative weight or coefficient, showing how it affects the dependent variable. This facilitates a model-building approach rather than a confirmatory/disconfirmatory one. Multiple regression methods can also be developed to allow the examination of residuals to improve the model.

In the work reported here, the use of the ANOVA revealed that matchtype was an important determiner of reading time. Regression was used to model matchtype in terms of 'load variables', essentially interactions between matchtype and sentence. The way in which the load variables were designed allowed the regression weights to be viewed as indicating how the cognitive load was broken down into separate components depending on the present and past load of matching, mismatching and neutral vocabulary items.

2.2.5 Conclusions for the Antonymy Experiment

The results of this experiment successfully replicate the finding of the semantic ordinal effect in Stenning (1986). The use of the counting task appears to have lengthened reading times but not to have altered the basic shape of the reading time curve. Different recall tasks and the use of non-binary material affect the reading times in interpretable ways but can all be accounted for by the same form of statistical model. In other words, it would appear that these manipulations have had little effect on the basic processes used but have caused a few differences in strategy to be made.

The modelling method itself has proved to be very successful. The observation that matchtype was important coupled with the exploratory modelling allowed by multiple linear regression has borne fruit. The same variable definitions in different linear models can account for all four quarters of the data. Much of the variance in the data is accounted for by variables that can be interpreted as independent processes acting on different components of the structures of the description. The way in which local processes accompanying mismatch detection are affected by the binary/non-binary distinction but not the menu/free recall distinction is revealed. The general strategy of delaying the recurrent representational processes until later on in the text for the harder tasks is also revealed by the regression modelling.

2.3 The Replication Experiment

The experiment described in this section consisted of a large data collection exercise designed to test various hypotheses and collect enough reading time and recall error data to produce 'firm' statistical models. The reading time data was modelled using the same multiple regression methods described in Section 2.2. The analysis of the recall error data is described in Chapter 3. The reading time model is reported fully in Stenning et. al (1988).

The experiment was designed to replicate one of the main findings of Stenning

(1986) — the observation of the semantic ordinal effect (SOE) (see Chapter 1). Three different determinate modes were used to determine that the SOE did not depend on a particular ordering of attributions. Two different vocabulary sets were used, one of people and one of geometric objects to check whether the semantic domain used made any difference to reading time. The people vocabulary set was split into two halves, one containing one syllable words and the other containing words of two syllables. This allowed an estimation of the extent to which the increasing reading times could be explained by articulatory rehearsal.

2.3.1 Experimental Details

Subjects

24 psychology students participated as part of a course requirement.

Materials

Subjects read texts of up to eight declarative sentences. The texts described two individuals (either people or objects) in one of three possible modes — Property by Property (PxP), Individual by Individual (IxI) or Multiple Attribution (MA). The multiple attribution texts are unique to this particular experiment and consist of two sentences, each describing one of the individuals in natural adjective order, e.g:

There is a strong young French nurse. There is a strong old Greek chef.

There were equal numbers of all eight matchtypes. The texts in the People vocabulary set were made up entirely of one syllable words or entirely of two syllable words (see Appendix A for vocabulary sets).

After the texts had been read the subject answered two questions. The questions took the form of an introducing noun paired with an adjective (e.g: "Is there a

Greek nurse?"). The answer to the questions was equally often "yes" or "no". The noun was equally often picked from the first presented individual as it was from the second presented individual. The dimension asked about was equally often matched or mismatched. Within these constraints the properties used in the questions were chosen at random.

Procedure

The experimental procedure was similar to the experiment reported above. The subjects read the text, sentence by sentence on the screen of a networked BBC microcomputer. Each text was preceded by a 'setting' that displayed the vocabulary dimensions on which the individuals would be described. The subjects were told to take as much time as they needed to be accurate in their recall. After they had finished reading, the subjects were required to give Yes/No answers to two questions. They then typed their recall using a simple menu. Finally, they were given a feedback sentence containing the correct description of the two individuals.

2.3.2 Summary of Basic Results

This section will concentrate on the way that the multiple regression model was developed for this data. The other results will be summarised. A full description of the results can be found in Stenning et al. (1988).

Vocabulary set was found to have no significant effect on the reading times. Although practice did have an effect (first half reading times being significantly slower than second half reading times), it did not interact with any other factors and so the data was collapsed over first and second halves.

For the texts made up from the people vocabulary set which were the only ones where the number of syllables in a vocabulary item were controlled, one syllable texts were read faster than two syllable texts. This difference in reading time did not increase regularly as the texts were read as would be the case if the semantic

	Individual 1				Individual 2			
Property:	A	B	C	D	A	B	C	D
IxI mode	1.36	1.57	1.82	3.16	1.63	1.74	1.94	2.45
PxP mode	1.39	1.62	2.04	2.56	1.36	1.95	2.36	2.73
MA mode	1st individual: 6.22				2nd individual: 4.80			

Table 2.9: Mean Reading Times (sec.) as a function of Individual, Property and Text Mode (both vocabulary sets)

ordinal effect was based on articulatory rehearsal. However there were large differences in reading times between 'short' and 'long' description for both sentences of the MA texts and sentence 4 of the IxI texts. This is good evidence of articulatory rehearsal of single individuals at these positions. It is clear that, although not able to account for the semantic ordinal effect, rehearsal is an important maintenance process for these texts. Further subtle analysis of the use of articulatory rehearsal in the processing of this type of text can be found in Stenning, Patel and Levy (1987) and Patel (forthcoming).

Table 2.9 summarises the main reading time results. The data replicates Stenning (1986) in showing that the semantic ordinal effect is stable over different vocabulary sets, modes and levels of practice. The fact that the reading time curves for IxI and PxP are so similar when both are sorted into property order is clear evidence that the reading time for a sentence increases as more is known about the individual referred to by the sentence. This conclusion is strengthened by other work using an even wider variety of modes of presentation (see Stenning, Patel and Levy 1987 and Patel forthcoming).

The difference in reading times between the two sentences in the MA mode was not significant. The mean reading time for individuals in the MA mode (5.5 sec) lies halfway between the reading times for the final sentence about an individual in an IxI text (2.8 sec)and the accumulated reading times of all four sentences in an IxI text (7.8 sec). Stenning et al. consider this evidence that these reading times reflect considerable constructive load as opposed to maintenance load because if maintenance load was dominant it would be expected that the MA sentence times

would be very close to the reading times for the final sentence about individuals in the other modes.

Clark (1973) has pointed out that treating language as a fixed effect may not allow experimental results to be generalised to different materials. He recommends that separate analyses considering subjects and materials as random variables should be combined to give F' or $minF'$ statistics for each main effect and interaction in an analysis of variance. There are many psycholinguistic experiments where relatively few test materials are selected by the experimenter, and it is wise to design such experiments so that a materials analysis is possible. In the MIT, however, each subject sees a different and very small fraction of the possible combinations of vocabulary items. These combinations are *randomly* chosen — in this experiment a subject saw 24 out of a possible (352,512) descriptions. This random selection of vocabulary dimensions weakens the need for a materials analysis because it ensures that a large and unbiased sample of materials are used. Even if we changed the experimental paradigm to make a materials analysis more practical it could only show that our effects will generalise to other combinations of the same vocabulary set The fact that there was no difference between the 'people' and 'object' vocabulary sets is reassuring in this respect. So, despite the fact that we cannot statistically demonstrate that our results generalise to all domains, we can be fairly sure that we can account for two representative domains. This is not to say that there are not subtle effects of vocabulary and, indeed, we make appeal to the use of associations based on personal past experience in our model of subjects' performance. However, the difference between vocabulary items appears to have little effect on the large differences in reading times observed here.

It is well known that the frequency of word usage can correlate with various psychological measures (see Henderson 1987, Morton 1969). To avoid complicating and weakening the power of the design, it was decided to ignore word frequency in selecting the vocabulary for the experimental texts. This decision has subsequently been vindicated by an experiment done by our research group that did systematically vary word frequency as part of the experimental design. There was a marginally insignificant main effect of this variable on reading time, it was small

(roughly 2%) and there were no interactions with other experimental variables.

2.3.3 Regression Model

The reading time data was modelled in exactly the same way as described in Section 2.2.3. However, a new variable was found to be necessary. In IxI texts it is clear that reading time increases as the first individual is described. This cannot be accounted for using MISLOAD, MATLOAD or LOCMIS since these are all zero during the first individual in IxI texts. Intuitively, it seems reasonable that a cognitive load is imposed by properties of an individual that cannot be assigned a matching or mismatching status with respect to the other individual. This load can be defined by a variable, NEUTLOAD, that is equal to the number of properties that are unresolved in this manner. In an IxI text it takes values of 1, 2, 3 and 4 for the first individual and then drops to 0 for all the sentences describing the second individual. For PxP texts it alternates between values of 1 for sentences describing the first individual and values of 0 for the sentences describing the second individual that can be directly compared to the previous sentence. The definition of NEUTLOAD is summarised in Table 2.4.

Separate models were built for IxI and PxP texts but they were so similar that the data was pooled to produce a single general model. The model is summarised in Table 2.10. The variance partition for the model is shown in Table 2.11. The observed and predicted reading times for the different matchtypes are shown in Figure 2.6

The regression model chosen by the statistical package is unsurprising. All the load variables gradually increase except for the last level of NEUTLOAD, for which there is a tripling between NEUT3 and NEUT4. This is at least partly due to the significant amount of articulatory rehearsal that takes place at sentence 4 of IxI texts. Apart from NEUT4, the recurrent variables are roughly ordered NEUTLOAD, MATLOAD, MISLOAD in terms of their contributions to reading time. LOCMIS contributes less to reading time than any other variable. The fact that the model fits so well for both text modes is a very successful result, indicating

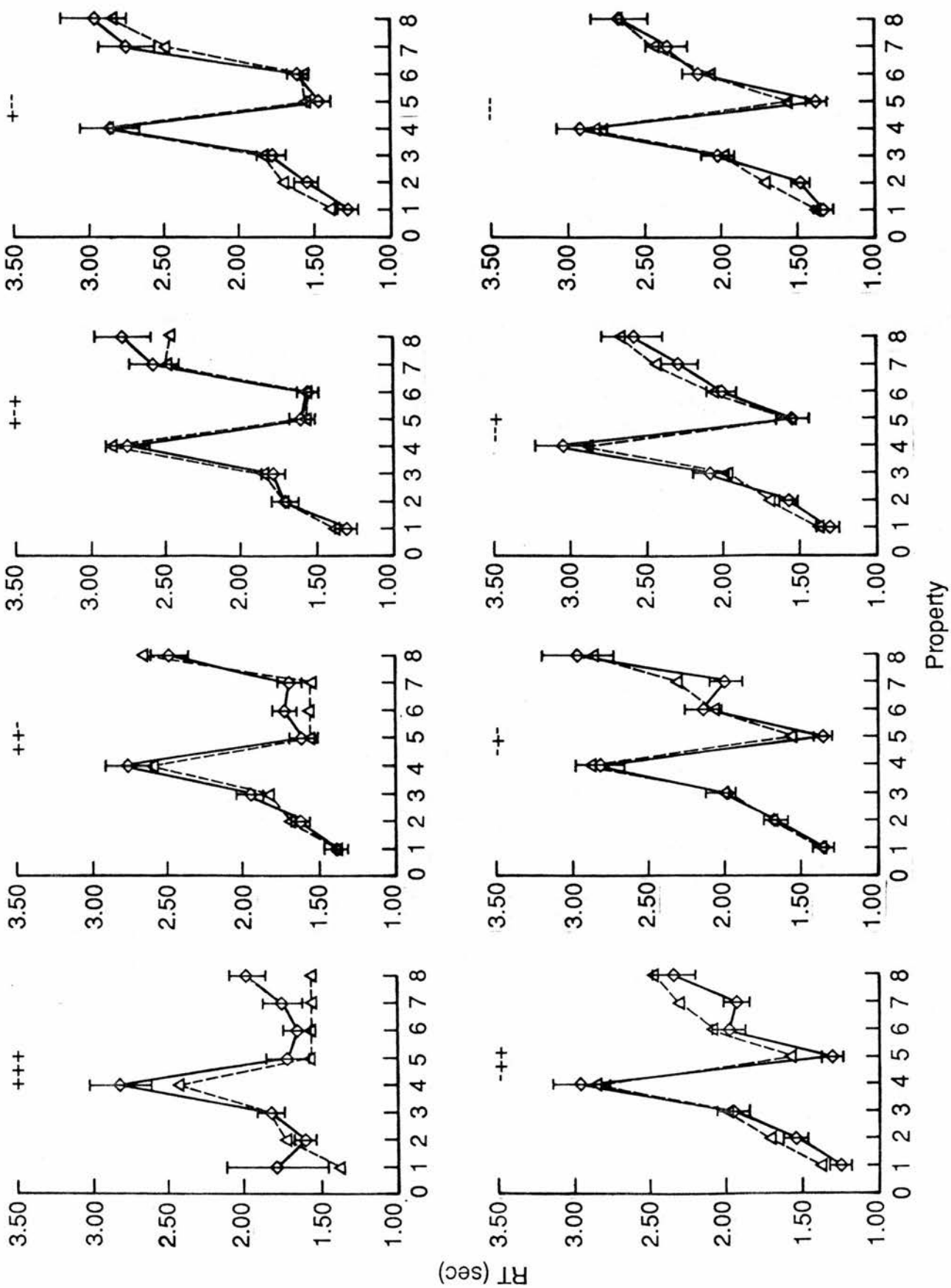


Figure 2.6: Observed and Predicted Reading Times for the Replication Experiment

Variable	Coefficient(sec.)	Standard Error
Intercept	1.09	0.066
NEUT1	0.28	0.049
NEUT2	0.47	0.092
NEUT3	0.73	0.092
NEUT4	2.07	0.092
MAT1	0.42	0.063
MAT2	0.58	0.090
MIS1	0.46	0.062
MIS2	0.79	0.076
MIS3	1.15	0.090
MIS4	1.37	0.150
LOCMIS	0.19	0.065

Table 2.10: Summary of the Regression Model for the Replication Experiment Reading Time Data

Partition	% Variance
Pure Error	87.66
Regression	11.89
Lack of Fit	0.45
'Success'	96.35

Table 2.11: Variance Partition for the Replication Experiment

that the processes are affected by the semantic structure of the descriptions rather than the surface order.

2.3.4 Comparison with the regression models for the Antonymy Experiment

The most significant difference between the models for the data of the two experiments is the introduction of NEUTLOAD. NEUTLOAD is less important for the PxP texts than for the IxI texts of the Replication Experiment but nevertheless is significant. The reason it is part of the model for the PxP texts in the replication data and not for the antonymy data is probably due to a slight difference in strategy because subjects in the former experiment have to deal with both IxI and PxP descriptions, and NEUTLOAD reflects vital processing during the first individual of IxI texts.

The general smaller size of all coefficients, shallower growth of recurrent load variables and smaller relative contribution of LOCMIS reflects the much easier time that subjects in the Replication Experiment had than the poor souls who had to endure the Antonymy Experiment. The counting task has a large effect on reading times, forcing subjects to take more care. Evidently it has a greater disrupting effect than the question answering task. The non-binary materials make the representational task much harder and force a strategy where reading times are large and processing tends to be partially delayed until the end of the text. Despite these changes in strategies, both models are surprisingly similar and, apart from NEUTLOAD, use exactly the same variable definitions to give very good accounts of the data.

The use of multiple regression modelling has proven to be very successful. The method has proved general across different experiments, text modes, recall tasks and methods of generating descriptions. Importantly, it has gone some way in explaining the origins of the semantic ordinal effect by providing a model of how the cognitive resources needed for the incremental interpretation of this type of description are partitioned according to different aspects of match/mismatch struc-

ture.

2.4 Summary of Reading Time Model

This section will summarise the hypotheses of the reading time model, contrast it with other work and discuss some of its predictions.

The reading time model embodies several hypotheses. First, it proposes that the semantic ordinal effect is due to the increase in cognitive load with the amount of information already known about the referenced individual. We propose that this load is ultimately caused by a search for associations between the incoming information and background knowledge that serve to bind properties together. The complexity of the search for associations is affected by the pattern of matching and mismatching in the description. This is because the pattern of matches and mismatches controls the difficulty of remembering which property is attributed to which individual — the greater the number of mismatches the harder the task.

Although the model contains the claim that the semantic ordinal effect is not caused by articulatory rehearsal, it does not preclude rehearsal as a component of mechanisms responsible for the performance of the task. Indeed there is some evidence that rehearsal can be important for more complicated descriptions (see Section 6.4).

The model is more fine-grained then many other models employing reading time methodology such as work by Sanford, Garrod and Kieras (see Chapter 2). The sentences in the MIT are pared to the minimum. Consequently the model could perhaps be viewed as a bridge between models springing from word list methodology and those using more naturalistic sentences.

The model predicts that the semantic ordinal effect should occur in descriptions with more complicated text modes than the ones used here. The effect is found in other experiments done within our research group (see Stenning, Patel and Levy, 1987; Patel forthcoming). The regression models fitted to the data from these

other experiments have similar load variables to the ones described here as well as more complicated ones connected with referential change. The reading time model also predicts that the structure of matching and mismatching of a description will be important in any model of the underlying representation supporting successful recall. This is shown to be the case in Chapter 3 where features representing whether a dimension matches or mismatches are important parts of a model that accounts for recall error frequencies.

2.5 Implications for the Representation

One of the strengths of the MIT is that it gives two windows on the representation of individuals and their properties: data on the incremental interpretation of the texts and data on structure of the underlying representation as revealed by errors in recall. It is important that these two views are not independent but feed each other and ultimately contribute to a unified model of how the attribute binding problem is solved for the material in this experimental paradigm.

The regression models of reading time data described in this chapter provide a springboard for the analysis of recall errors described in chapter 3. The most important clue is the importance of inter-individual relations, specifically the matching and mismatching across vocabulary dimensions, in the construction processes for the representation of the information in the descriptions. Bearing this in mind, we now go on to describe the methods used for modelling recall errors.

Chapter 3

Recall Error Modelling

3.1 Introduction

This chapter describes the development of a model of memory for the descriptions of individuals that was based on the statistical modelling of recall errors from the MIT. The history of the project is traced from the initial inspiration for the model (that came from the examination of the common classes of errors) to the way in which multiple regression techniques were used to extract two different types of model. The process of refining a model using these techniques is examined. The theoretical implications of the final model are discussed. Finally, the reasons for wanting to extend the modelling framework to one closer to a process model are discussed.

3.2 What Are We Seeking to Model?

The MIT is designed to pose a severe attribute binding problem for subjects. They are presented with many descriptions constructed from a restricted vocabulary set and the pair of individuals described in each text vary across binary dimensions, the dimensions matching or mismatching unpredictably. One would expect these loads to make the task of assigning the correct property to the correct individ-

ual very taxing and to produce some errors when recall is tested. In such an experimental paradigm it is hoped that the structure of the error data collected will help the experimenter gain some insight into the underlying representation in memory that supports the task required of the subject. In this case, the task is to remember the correct assignment of each property to each individual. Since we know that the knowledge a subject has supports her memory (e.g. Miller 1956, Bransford et al. 1972) we would expect any model of attribute binding to allow for the interaction of knowledge from the rest of memory with the construction of a new representation.

Any model of this data must be able to explain:

1. How the properties of an individual are 'bound' to that individual.
2. How separate individuals can be individuated when the subject is asked to recall them.
3. How the underlying representation gives rise to the observed errors when disrupted in some way.
4. How knowledge can be brought to bear on the binding of properties to individuals. This requirement precludes the vacuous model of binding as a content-less link between tokens representing properties and a token representing the 'name' of the individual.
5. Any model of the characteristics of the representation of individuals should be compatible with what is known about the cognitive loads imposed during the incremental interpretation of the information. In other words, a successful model should shed further light on the time-consuming processes successfully described by the reading time regression models.

The actual process of modelling started with an attempt to infer some of the main properties of the underlying representation from the gross properties of the error data. The next stage was to classify the *patterns* of errors and make the statistical analysis of the frequencies of these error pattern categories the goal.

3.3 Implications of the Reading Time Models

One of the aims of the MIT is to use both reading time *and* recall errors as probes on human representation. It would be disappointing then, if the regression models of reading time processes did not give us a head start on how to model recall errors. This did prove to be the case. The interpretation put on the semantic ordinal effect (see Chapter 1) was that, as subjects read the material, they made some sort of semantic effort that got harder as more information was input. From what has been said above about the use of content in human memory, it would seem natural to suppose that this semantic effort was relating the incoming information with past knowledge in memory. We claim (see Chapter 2) that this 'relating' was done to 'recruit' associations to give mnemonic support to the incoming material. The mnemonic support would allow the subject to make the correct distinctions between groups of properties that are necessary for correct recall of two separate individuals from the recall menu. One of the major factors in the reading time regression model was the immediate and past history of matching relations between properties across dimensions. It would seem likely that this relation would also be important in the resulting representation.

3.4 The Error Data

The data described here is from the Replication Experiment. This data is described in (Stenning, Shepherd and Levy 1988) as is the first regression model. The second and more fully developed model is described here for the first time, a previous version of it appearing in (Stenning and Levy 1988). The purpose of repeating the description of the original model is to describe the development of the project and to enable the two models to be compared.

17 subjects produced data from the recall of 1537 texts. There was a weak speed-accuracy trade-off (the correlation between accuracy of the whole paragraph and total reading time of the paragraph was -0.2). It was shown that this trade-off

was mostly global — a fast reading time for one sentence affected the recall of the whole paragraph and not just the material in that sentence. The only local effect was a difference of 0.99 sec in reading time for sentences recalled correctly and sentences recalled wrongly for the final property of the first individual.

At this early point it is worth describing and making some comment on the way that the recall data is scored in our experiments. We are only concerned here with descriptions of pairs of individuals. For tasks where the order of recall of individuals is not cued, the subject is presented with a menu and asked to recall 'one of the individuals' . When this recall has been completed an identical menu is presented with the instruction 'Recall the other individual'. When scoring the recall, some decision must be made for the assignment of a particular recalled individual to a corresponding presented individual. This decision is made by a simple best-fit algorithm which makes the assignment that yields least recall errors. If there is a tie between the two possible assignments then this is broken by assigning the individuals in the order in which they were presented. This tie breaking device is justified by the finding that unambiguous recalls are usually made in presented order. The allowing of unrestricted recall order and having to use a best fit method to ascertain the correspondence between recalled individuals and presented individuals is done so as not to impose any unnecessary constraints on the processes used by subjects in their recall. Other experiments (Stenning, Patel and Levy 1988, Patel forthcoming) have cued the recall order, concluding that the fact that the second recalled individual has poorer recall is due to the interfering effect of recalling the first individual rather than a strategy of recalling the best remembered individual first.

Tables 1, 2 and 3 from Stenning, Shepherd and Levy (1988) are reproduced here to make the description of the gross error characteristics clear before going on to discuss the statistical modelling methods used to account for the error patterns hidden in this data.

Despite the taxing nature of the experimental task, recall was very accurate. The mean number of errors was 0.56, standard deviation 1.0, out of the 8 properties.

	Single error		Multiple error	
R-Individual 1				
Stimulus position:	First	Second	First	Second
	16.0	14.0	2.1	3.6
R-Individuals 2				
Stimulus position:	First	Second	First	Second
	16.0	18.0	8.4	8.1

Table 3.1: Percentage of single and multiple errors as a function of stimulus position and order of recall (N = 1537)

71% of the paragraphs were error free. There was no evidence of practice or proactive interference effects — the mean unit error was 0.53 in the first half and 0.59 in the second. After applying the best-fit recall scoring algorithm, it was determined that 70% of the paragraphs were recalled in stimulus order and 30% in reverse order.

Examination of Table 3.1 reveals that recall errors were more correlated with recall order than presentation order — there were slightly more single errors, and relatively many more multiple errors, on the second recalled individual than on the first recalled individual. As mentioned above in the discussion of the recall algorithm, other experiments where recall was cued suggest that this observation is a result of the recall of the first individual interfering with the recall of the second.

The most interesting fact from this data is the relatively large number of multiple errors — more than would be expected from the frequency of single errors. This leads us to believe that the errors are reflecting some underlying representational structure, i.e. errors on the different properties are not independent of each other (see Jones 1978) and must therefore be bound together in some structure in memory. This observation is probably the most important factor of the data — it allows us to claim that analysis of the types and frequencies of the different errors that have been made can go some way to determining the structure of the underlying representation because they reflect dependencies between the representations of the different properties.

	Property A		Properties B–D	
	Matched	Mismatched	Matched	Mismatched
1st individual	–	0.72	1.55	2.45
2nd individual	–	1.56	6.79	6.40
Both individuals	–	4.75	1.85	4.42

Table 3.2: Percentage of first, second, and both individual errors as a function of matched and mismatched properties

	Property			
	A	B	C	D
Single errors				
R-Individual 1	4.68	2.99	3.51	4.03
R-Individual 2	3.71	4.03	5.27	4.36
Multiple errors				
R-Individual 1	0.78	1.24	1.63	2.02
R-Individual 2	2.67	4.55	5.14	5.66

Table 3.3: Percentages of single and multiple errors across properties within individuals

Table 3.2 only strengthens this evidence that the data will prove rewarding to analyse. Here we see that there is a correlation between errors on both individuals for all dimensions and that this is strengthened for mismatching dimensions. The correlation is particularly strong for the introducing dimension, presumably because subjects know that this dimension is always mismatched and are unlikely to make a single error and recall it as a matched dimension (this only happened 35 times). It is much more likely that the introducing dimension will have a double error, reflecting a mistaken assignment of properties. It is clear from Table 3.2 that the recall of a property is strongly affected by whether it matches or mismatches with the corresponding property of the other individual. Also, the special status of the introducing property is reflected in its relatively low susceptibility to error and high correlation between errors on both individuals.

Table 3.3 tabulates the data for errors within individuals. Single errors are fairly flat across the four properties and there are a few more on the second recalled individual.

Response Type	Response							
Correct	tall	happy	Polish	bishop	short	happy	Swiss	dentist
Single	**short**	happy	Polish	bishop	short	happy	Swiss	dentist
Individual Polarity	**short**	happy	Polish	bishop	**tall**	happy	Swiss	dentist
Property Polarity	tall	**sad**	Polish	bishop	short	**sad**	Swiss	dentist
Double Complementary	tall	**sad**	Polish	bishop	short	happy	**Polish**	dentist
Double Homogeneous	**short**	happy	**Swiss**	bishop	short	happy	Swiss	dentist

Table 3.4: Common error categories

There is a trend of the tendency of a dimension to participate in multiple errors running from dimension A to dimension D. There is a clearly greater number of multiple errors on the second recalled individual. The log-linear models reported in Stenning, Shepherd and Levy (1988) show a tendency for property D to be correlated with the fate of other properties.

The correlations observed between the fates of properties across dimensions and within individuals is strong evidence that the recall error data is reflecting some underlying dependent structure. This is good news because it means that we have a chance of inferring something about the representation of individuals from observable error data.

The next stage on the way towards modelling the representational structure revealed by recall errors was to develop a classification of the error data that would allow it to be modelled statistically.

3.5 Evidence For Redundancy in Simple Patterns of Errors

Single errors are the most common type of recall error, but as explained above, there were more multiple errors than would be expected from the frequency of single errors if individual properties were represented independently. Important clues about the nature of the underlying dependencies between properties was gained by examining the common types of multiple error. Some of the common response classes are displayed in Table 3.4.

The most common response type was, of course, a *correct* response (70.7%).

We distinguished between four categories of single error, depending on whether the error was on the first or second recalled individual and whether it was on an originally matching or mismatching dimension.

The simplest categories of multiple error are the double errors on single dimensions (see Table 3.2). These are more frequent on mismatched dimensions than matched ones. An example of a double error on a mismatched dimension would be recalling a tall thin Swiss dentist and a short fat Polish bishop as a tall thin Polish dentist and a short fat Swiss bishop. This can be seen as remembering that the nationalities are different but forgetting which nationality to assign to which individual. We call this sort of error an *individual polarity error*. It would seem natural to model the occurrence of this sort of error as the remembering of one 'item' of information (the fact that the nationalities were different) while forgetting one or more others (whatever supports the assignment of nationality to individual). The gross correlations between the fate of properties within individuals and across property dimensions suggests that the supporting representation contains dependencies between the properties of an individual as well as between the corresponding properties of the pair of individuals. The character of the individual polarity error suggests that the underlying representation is redundant, since it appears that one of these types of information is preserved while the other is missing. This suggests a degree of overspecification of the information to be remembered for cases when no errors are made.

A much less common error is what we call a *property polarity error*, a double error arising from recalling the wrong property for a matching dimension. This could be explained in a similar way to the individual polarity error but the final model should be able to account for the fact that it is less frequent than a property polarity error.

Further sub-divisions of the error data were made by grouping together the triple errors resulting from combinations of individual and property polarity errors with single errors. The individual polarity errors were divided into four sub-types by

distinguishing between whether the singleton was on the first or second recalled individual and whether it was on a matched or mismatched dimension. These subdivisions were not made for property polarity errors because of sparsity of data.

The remaining double errors were ones where there were single errors on two different dimensions. These were divided into two main headings — *double complementary* and *double homogeneous*. Double complementary errors are ones where one originally matched dimension is recalled as a mismatched one and another, originally mismatched dimension, is recalled as a matched one. Double homogeneous errors are one where two matching or two mismatching dimensions are recalled as two mismatching and two matching dimensions respectively. Both of these types are split into three subcategories depending upon whether both errors are on the first individual, both errors are on the second individual or both individuals have one error.

A curious error type found was one where the structure of matches and mismatches along the dimensions seemed to have been reversed (apart from the mismatching introducer). This type of triple error was called a *mirror* error.

The other errors observed did not appear to fall into any large groups and were all fairly severe errors. All of these other errors were grouped into a 'miscellaneous' category.

Table 3.5 lists all the response categories along with their observed and chance probabilities.

3.6 Recall Data from the Antonymy Experiment

Although it produced enough reading time data for regression modelling, there was not enough recall error data from the antonymy experiment for a full analysis. Some interesting results did emerge however. The following results are fairly qual-

Abbreviation	Response type	Observed	Chance
corr	Correct	0.707	0.006
misc	Miscellaneous	0.014	0.569
sg1+	Single error on R-1 matched	0.018	0.009
sg1-	Single error on R-1 mismatched	0.025	0.015
sg2+	Single error on R-2 matched	0.024	0.009
sg2-	Single error on R-2 mismatched	0.044	0.015
ipol	Individual polarity error	0.066	0.015
is1+	Individual polarity with sg1+	0.005	0.016
is1-	Individual polarity with sg1-	0.004	0.023
is2+	Individual polarity with sg2+	0.012	0.016
is2-	Individual polarity with sg2-	0.008	0.023
2cs1	Double complementary both on R-1	0.008	0.019
2cs2	Double complementary both on R-2	0.008	0.019
2cdf	Double complementary on R-1 and R-2	0.014	0.032
dhs1	Double homogeneous both on R-1	0.004	0.019
dhs2	Double homogeneous both on R-2	0.007	0.019
dhdf	Double homogeneous on R-1 and R-2	0.002	0.037
ppol	Property polarity error	0.012	0.009
pp+s	Property polarity with single	0.005	0.055
mirr	Mirror image matching structure	0.008	0.049

Table 3.5: Observed and chance probabilities of occurrence of response categories

itative and some might be fruitfully investigated in a further larger experiment.

The different variable delays (5, 10 or 20 seconds) filled with counting backwards in threes had no differential effect on the number or distribution of errors. There was a slight tendency for the ratio of recalls in presented order to those in swapped order to increase as the delay increased.

For both menu and free recall there were slightly more errors for non-binary texts than for binary ones. The non-binary texts contain more information since their vocabulary pairings are more unpredictable. It is perhaps surprising that the difference between binary and non-binary texts is not greater, especially for free recall where there is no cueing. What is perhaps happening is that the longer reading times required for the non-binary texts build a representation that is almost equally as robust as that for a binary description. Subjects appear to be taking as long to read as necessary to build a 'satisfactory' representation. This might also explain the fact that the gross number of errors in the antonymy experiment is about the same as that for the replication experiment even though the reading times are much longer.

Unlike the replication experiment, there is no clear trend in multiple errors across the different vocabulary dimensions. Another aspect of the replication data that is not found in the antonymy experiment data is the greater number of errors on the second presented individual.

As well as sparsity, the free recall data from the antonymy experiment is especially hard to analyse because of the complications of omissions and intrusions (vocabulary items from previous texts) that inevitably occur. There is a slight tendency for a greater number of intrusion errors in the non-binary texts. This is unsurprising because if a subject forgets an item in a binary text she can generate it from its antonym or associate.

It is difficult to compare the frequencies of different recall error categories because of the small amount of data. However, there are multiple errors including the important polarity errors in the antonymy experiment data.

3.7 Theoretical Assumptions Underlying the Statistical Modelling

The characteristics of the gross errors and the distribution of error patterns gave some fairly good evidence for aspects of the structure of the underlying representation. There is good evidence for dependencies between properties, both within and between individuals. Aspects of this dependent structure appear to be independent and redundant. These observations support the following assumptions that lead to a framework whereby the frequencies of different error categories can be modelled statistically:

1. The more similar a possible response is to a stimulus, the more likely it is to be 'confused' with the stimulus and recalled as a result. Thus, the most likely response will be the correct one. Of course, this assumes we have a conception of the representational similarity between pairs of descriptions.
2. The underlying representation is based on independent features that provide the dependent structure observed between different properties by the correlations in the error data. So, these features will link the fate of different properties within an individual (intra-individual features) as well as properties across a dimension (matching features). Every description is represented by the values of all of the individual features. The features are named in terms of dimension names and stimulus individuals that are involved. The values taken by the features differ between the two main models and this is discussed below.
3. There is considerable overspecification or redundancy in the specification of the information involved by the set of features in the representation.
4. The probability of a particular class of error will be predicted from the degree of feature disruption it causes. This will depend on the number of features disrupted and their relative importance. The greater the degree of disruption the smaller the chance of that particular type of error occurring since the more disruption the less similar the stimulus and the response.

3.8 The Statistics Used

The previous assumptions provide the basis of a linear model to account for the different classes of error frequencies. The dependent variable is a measure of the frequencies of different error classes and the independent variables will provide a measure of similarity between stimulus and response based on the feature values that they share. Multiple linear regression allows the extraction of a linear model from this data based on the above assumptions. It allows each of the features to be given a different weighting (coefficient) in the measure of similarity. Multiple regression is an attractive technique because various procedures exist (Draper and Smith 1981; Dixon et al. 1983) that allow a number of candidate predictor variables to be examined and a 'best' model to be picked. This feature allows the technique to be used for *modelling* rather than simple data description, since it allows us to determine the percentage of variance accounted for by different sets of independent variables.

The error frequency data was very skewed since most of the responses were correct ones. The chance probabilities of particular error classes differed greatly because of the nature of the task. For example there was only 1/136 chance of a correct response while there was a 57% chance of a 'miscellaneous' error (see Section 3.5). To provide a variable that reflected the frequency of an error class in proportion to the opportunity of making that error, the frequency of each error type was divided by the chance opportunity of making that error. This adjusted frequency was then logged to get a more normally distributed variable. The skew and kurtosis statistics are reduced from 4.2 and 20 for the raw variable to 0.7 and 5.3 for the logged and adjusted variable.

The random probability of each recall error category was calculated by assuming that there is an equal chance of making any of the possible menu responses for any given stimulus. Eight representative stimuli were chosen, one from each matchtype. Each stimulus was paired with each of the possible responses from a recall menu. These artificial 'stimulus-recall pairs' were scored in the same way as the human data and the number of occurrences of each error category was

counted. The chance probability for each type of error was then calculated by dividing its mean frequency by the number of possible responses from any single recall menu (136).

The independent (predictor) variables were the features. The value of a feature variable for a given response type was the mean proportion of times the feature was recalled correctly when that type of error was made. Thus all the independent variables have a value of 1.0 for correct responses since all features are always intact for this response category.

The actual data modelled were the log adjusted frequencies of each different error category for each of the eight matchtypes. Since 20 different response categories were chosen, there was a maximum of 160 possible data points. In actual fact, not all error categories occur for each matchtype.

The best possible subsets program use in the analysis of the reading time data (see Chapter 2) couldn't handle the number of variables required for this analysis so a stepwise regression procedure was used — the P2R program from the BMDP statistical package (Dixon et al. 1983). The task of the statistical package is to pick an equation specifying the set of feature variables and coefficients that best predicts recall performance. The coefficient can be interpreted as a degree of salience of the particular feature, specifying the *amount* of similarity contributed by two descriptions sharing that feature. The statistical procedure as a whole extracts a similarity metric, picking the most significant features and giving them weightings. The set of features and weightings amount to a model of the underlying representation.

3.9 The Process of Model Refinement

The particular statistical technique that was used allowed a large degree of choice in how to build a particular model. Since each possible description can have 136 possible responses and there is only a limited amount of data, it is necessary to

divide up the responses into different categories to avoid sparsity of data problems. The particular choice of error categories was made by examining the data to see what the common types of error were. A balance was struck between dividing the errors up into interesting categories and having enough occurrences of a particular category. For example, there are four categories of individual polarity with singleton but only one of property polarity with singleton because there were so few errors of this category made. It was also a matter of choice to divide the data up into frequencies of the different error types for each of the matchtypes. This was done because so many of the error categories were dependent on the matching or mismatching of a property dimension and so might be expected to differ in frequency for different matchtypes. To take an extreme example, response categories that depend on having a matching dimension can never occur for the fully mismatching matchtype 8.

The multiple regression procedure allowed several candidate supersets of feature variables to be tested. The variables could then be modified based on the success of the model and an examination of the residuals (see Chapter 2) as well as theoretical justifications. The main differences in the two models described here are based on theoretical considerations rather than pure data-fitting or variance capturing criteria.

As explained above, the existence of errors such as the individual property polarity error is strong evidence for an aspect of the underlying representation that encodes the matching or mismatching of a dimension. Both of the following models contain four matching features, one for each dimension, that take a value of 1 (or 'true') if the dimension matches and 0 (or 'false') if the dimension mismatches. This type of feature can account for the existence of individual polarity errors because it will be *preserved* when this type of error occurs, contributing a degree of similarity between stimulus and response and thus making the response more likely.

There is a general tendency for there to be more double complementary errors than double homogeneous errors. This can be modelled by the inclusion of the 'number of matches' (NMAT) feature which takes a value equal to the number of

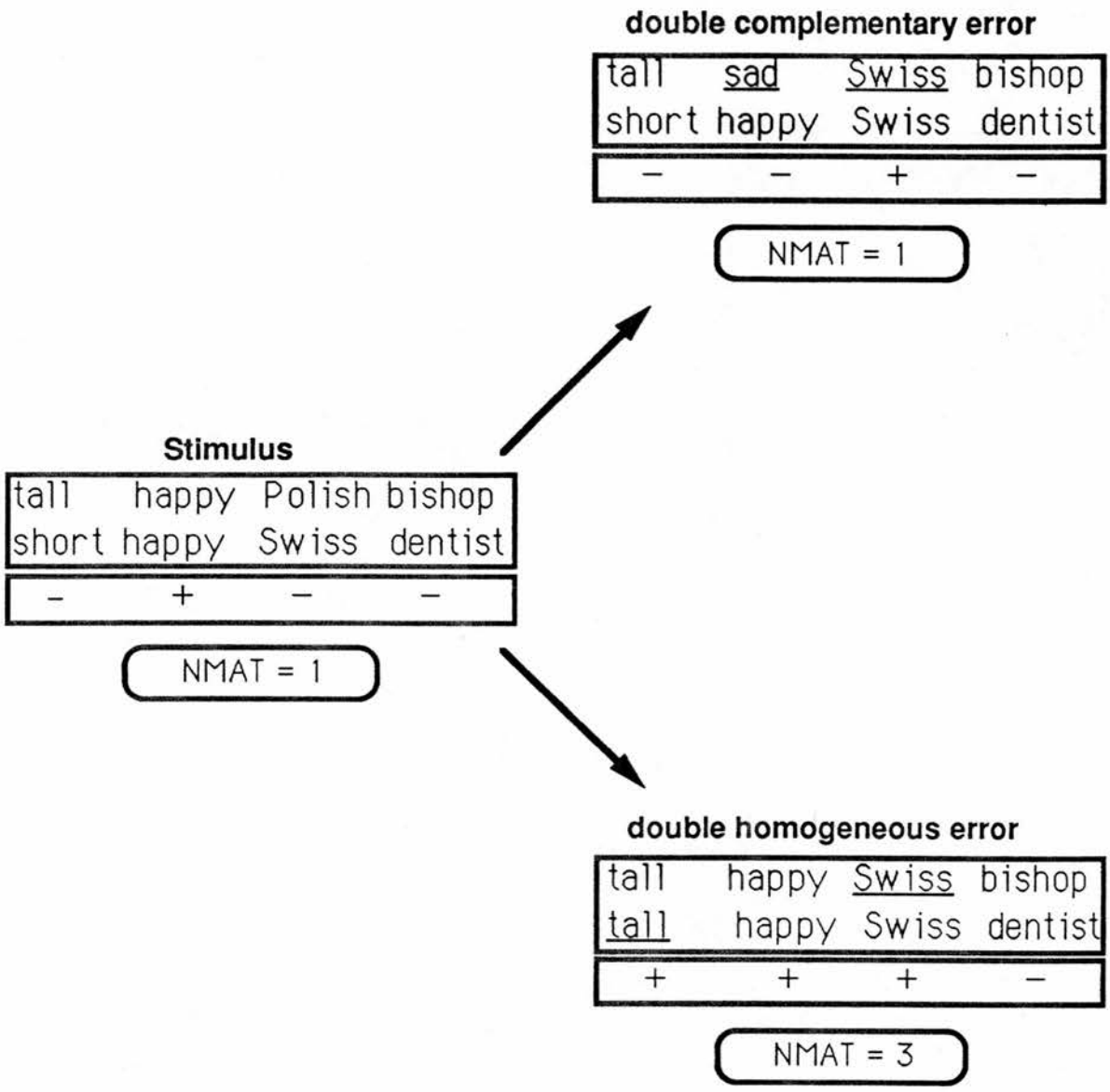
matching dimensions. This feature is preserved by double complementary errors but changed by a value of 2 for double homogeneous errors (see Figure 3.1). So, although both these errors are double errors on different dimensions, the extra degree of similarity conferred by the preservation of the nmat feature will tend to predict that double complementary errors will be more frequent.

One problem with some of the models picked by the multiple regression was that although explaining a large amount of variance, the set of variables picked did not distinguish all the possible descriptions. This became a problem when the statistical models were extended into PDP models (see Chapter 5, Table 5.4). This problem was fairly easily solved by minor intervention in the stepwise regression process (see Section 3.11). The problem only occurs for the 'instantiation model'.

3.10 The Tagged Model

I shall refer to the first model developed as the 'tagged' model because the intra-individual features are labelled as to whether they refer to the first or second individual. This model is reported in Stenning, Shepherd and Levy (1988) and Levy and Stenning (1988). For this class of model, a feature took a value that depended on the vocabulary of the properties that were tied together by that feature for a specified individual. For example, the feature BCD1 would take the value that depended on the vocabulary items for dimensions B, C and D for the first individual. The match features, DIMAMAT, DIMBMAT, DIMCMAT and DIMDMAT take values of 'matched' or 'mismatched'. NMAT takes values between 0 and 4 but when a stimulus feature set is compared with the feature set for a response, NMAT is scored as either 'preserved' or 'disrupted' like the other features. Making the comparison scalar, so that the numerical values of the two different NMAT features were compared resulted in a worse model. An example of the values taken by the features from the tagged model can be seen in Table 3.6.

One of the most notable points of the model is that all four match features are



Although both responses are double errors, NMAT changes for the double homogeneous error making it less likely to occur.

Figure 3.1: The value of NMAT for double complementary and double homogeneous errors

For the description: 1: short happy Swiss dentist 2: tall sad Swiss bishop	
Feature	Value
DIMAMAT	mismatch
DIMBMAT	match
DIMCMAT	mismatch
DIMDMAT	mismatch
NMAT	1
BD2	tall Swiss
AD1	short dentist
CB1	happy Swiss
AC2	tall bishop
B2	Swiss
BCD1	short happy Swiss

Table 3.6: Example values of the features in the tagged model

Feature	Coeff	S.E.	P(correct)
DIMAMAT	1.01	0.15	0.98
DIMBMAT	0.21	0.09	0.93
DIMCMAT	0.49	0.08	0.91
DIMDMAT	0.34	0.08	0.91
NMAT	0.23	0.07	0.83
BD2	0.69	0.11	0.86
AD1	0.37	0.09	0.87
CB1	0.35	0.13	0.90
AC2	0.48	0.08	0.85
B2	0.36	0.13	0.92
BCD1	0.67	0.14	0.85

adjusted R^2 = 0.85;df=11/105;Intercept = -3.24

Table 3.7: Summary of tagged feature regression model

present. The highest coefficient of the whole model is DIMAMAT — this is not surprising since this feature will always have the value 'matched' for a valid recall since the introducing dimension always mismatches, and subjects must soon learn this. Their recalls rarely disrupt this feature (its accuracy is 0.98 — see Table 3.7) because of its special status of only ever having one value for a well formed recall and so it confers a high degree of dissimilarity if not shared by two representations, thus making this eventuality relatively unlikely.

NMAT is included but has a relatively low coefficient — this is presumably because, despite its usefulness in distinguishing double complementary and double homogeneous errors, most other errors (apart from polarity errors) will disrupt nmat and so it is not particularly predictive of the frequency of any particular response type.

The intra-individual features are 'fully connected' for both individuals i.e. every property of both individuals is represented at least once. This means that the intra-individual features alone can represent the description. So, the presence of the match features and nmat represents a large degree of redundancy. In fact, there is even redundancy within the intra-individual features, e.g. B2 and BCD1 could be dispensed with and the remaining intra-individual features can still represent the original description adequately.

The overall fit of the model is very good. It accounts for 85% of the total variance in the data and there is undoubtedly some pure error that can't be accounted for although this is hard to measure. The fit to the 'contour' of the error categories (see Figure 3.2) is satisfying but there are some interesting inaccuracies in prediction. The overprediction of miscellaneous (usually very severe errors) and underprediction of correct responses would be expected for a redundant underlying system. The system will use this redundancy to *correct* errors and thus will have more corrects and less miscellaneous errors than can be accounted for by a simple linear model. The fact that the regression will not find all the redundant features because there will be some weaker ones contributing to error correction that will not reach significance may explain some of the lack of fit.

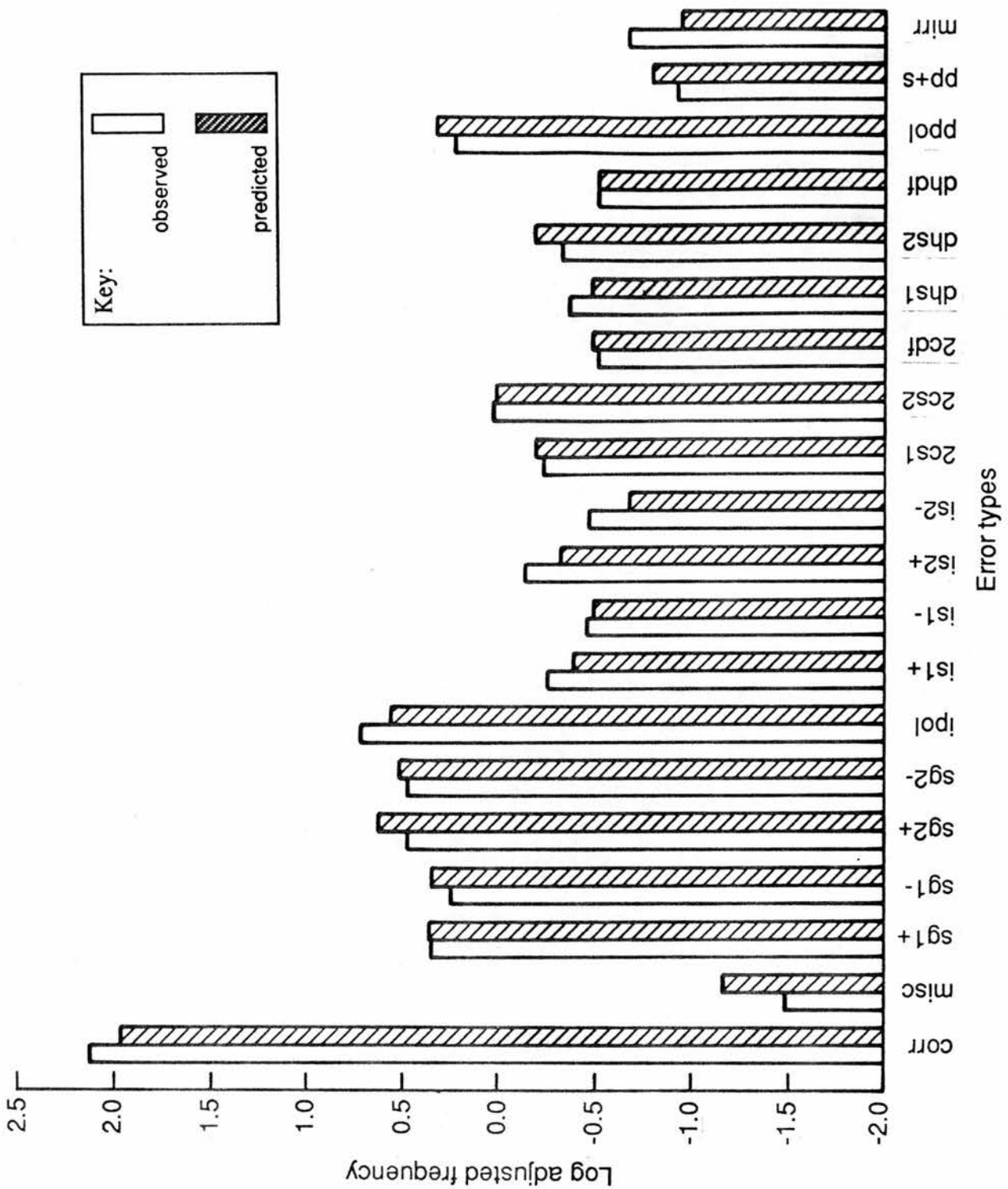


Figure 3.2: Histograms of Observed and Predicted Error Frequencies for the Tagged Model

As described in Stenning, Shepherd and Levy (1988), other combinations of intra-individual features fit the data almost as well. The important factor seems to be that models accounting for large amounts of variance have roughly the same number of features and the spread of sizes (number of properties included in a feature) is roughly the same. Interestingly, the best models do not appear to be centralised around any single property. It might have been expected that the introducing dimension would appear in most features to identify the individual concerned but this was not the case.

3.11 The Instantiation Model

Although a successful description of the error data, there are some criticisms we can make about the tagged model. Because of the way that the intra-individual features are labelled for the particular individual they refer to, the model begs the question of how a property is attributed to the correct individual — we know that the bishop is Polish if we know that individual 1 is a bishop and individual 1 is Polish. So, to some extent, the tagged model does not adequately explain how the attribute binding problem is solved because it does not identify a mechanism whereby a property can be 'tagged' as to which individual it refers to. This deficit lead us to find out whether a model that was based on features that were not tagged by individual could successfully account for the data. We called this the 'instantiation model' because it is based on features that either took a value of 'true' (were instantiated) or took a value of 'false'.

The distribution of recall errors observed suggests that the underlying representation is redundant. The matching and mismatching of dimensions appears to be important and it is also clear that the representation of intra-individual 'links' must also take place. Some of the errors also suggest that the number of matching and mismatching dimensions is represented. These observations led us to hypothesise separate features that describe the existence of a match or mismatch on any dimension, describe the existence of any of the possible combinations of vocabulary items within an individual and describe the number of dimensions that

match. In the 'tagged model', there is a feature for every possible combination of dimensions for the first individual and every possible combination of dimensions of the second individual. These features are labelled as referring to either the first or second individual and take the appropriate vocabulary items as values. In the 'instantiation model' these intra-individual features are not labelled by individual and the features are named according to the combination of vocabulary items they instantiate when they take the value of 1.

Candidate features were all possible single, pair, triple and quadruple combinations of the vocabulary items for a given menu. Match features were the same as the tagged model (except we now called their values 'true' and 'false' rather than 'matched' and 'mismatched'). Somewhat awkwardly, NMAT stayed the same too, taking values of between 0 and 4, although as before it was only ever scored as 'corrupted' or 'preserved' when a stimulus and response were compared. For this model, the representation of a description consists of the truth values for the feature set, NMAT taking a numerical value. We must now posit an inferential process to take place to identify which property should be attributed to which individual.

The model described here is similar to the one described in Stenning and Levy (1988). However, the features in that model, while accounting well for the variance in the data, did not logically distinguish all possible descriptions. This is unsatisfactory because it predicts that some descriptions will always be error prone. It also makes the extension of the model as a PDP network awkward (see Chapter 5). This was remedied by intervening in the stepwise regression process. The stepwise algorithm used (Dixon et al. 1983) calculates an F ratio for each variable and adds the variable with the highest F statistic to the model as long as the value is greater than 4.0. If the F ratio for a variable decreases below 4.0 at any stage the variable is removed from the model at the next step. Early in the regression reported in Stenning and Levy (1988) (after step 3) the variable with the highest F ratio was C (24.68) closely followed by DIMCMAT (22.13) and so C was added to the model. After C is included, the F ratio for DIMCMAT diminished and it was never included. The problem of logical incompleteness would be solved

Feature	Example	Coefficient	S.E.
D̃C	tall happy	0.29	0.11
DÃ	short bishop	0.26	0.10
CA	happy dentist	0.38	0.09
BA	Swiss dentist	0.40	0.12
D̃C̃B̃	tall sad Polish	0.76	0.10
D̃BA	tall Swiss dentist	0.38	0.10
D̃CA	tall happy dentist	0.24	0.11
C̃BÃ	sad Swiss bishop	0.52	0.10
DC̃B̃Ã	short happy Polish bishop	0.70	0.21
DC̃B̃A	short sad Polish dentist	0.50	0.12
DD/D̃D̃	both short or both tall	0.61	0.08
CC/C̃C̃	both happy or both sad	0.53	0.09
BB/B̃B̃	both Swiss or both Polish	0.24	0.09
AA/ÃÃ	both dentists or both bishops	1.10	0.16
nmat	more than one match	0.22	0.08

adjusted $R^2 = 0.84$; df = 15/101; Intercept = -5.16

Table 3.8: The Instantiation Regression Model

if DIMCMAT were included in the model. C is the only singleton feature, and does not easily fit with the rest of the model. If variable C is eliminated from the variable set, the regression picks a similar but logically complete model including DIMCMAT. The new regression model is summarised in Table 3.8. The features are named after the vocabulary items whose presence they denote. For a dimension X, one vocabulary item is denoted by 'X' and the other by 'not-X' or 'X̃'. For example, 'short happy' would be written as 'DC' and 'tall sad' as 'D̃C̃' Examples of fragments of a description that would make each feature true are given in Table 3.8. The observed and predicted log adjusted error frequencies are plotted in Figure 3.3.

A comparison between the tagged and the instantiation models reveals a few differences. Firstly, there are four additional intra-individual features in the instantiation model. The match features have similar coefficients apart from that for dimension D which has a higher coefficient in the instantiation model. NMAT has very similar coefficient in the two models. There are no uni-dimensional features in the instantiation model while the tagged model has one, and the instantia-

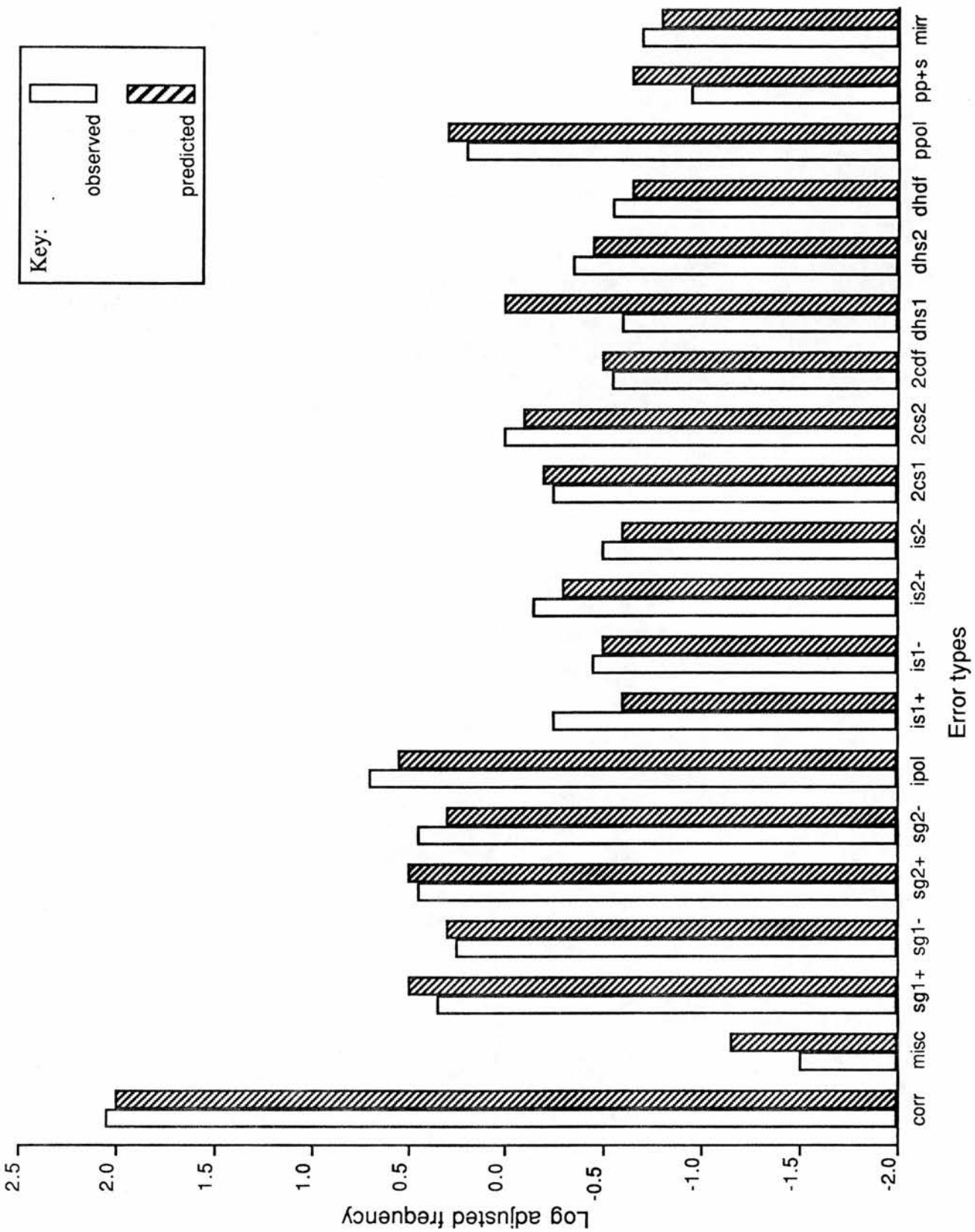


Figure 3.3: Histograms of Observed and Predicted Error Frequencies for the Instantiation Model

tion model has two four-dimensional features where the largest feature size in the tagged model is three. Like the tagged model, there is no evidence of centralisation around a particular vocabulary dimension. The models account for a similar proportion of the total variance.

3.12 Summary of recall model

The main claim of the recall model is that the underlying representation involved in the performance of the MIT is made up from distributed and redundant 'features'. These features are contentful associations that serve to bind together several of the vocabulary items in a description. Differences in the material to be remembered may make the search for associations harder and this is reflected in reading times. For example, a greater effort needs to be expended to distinguish two individuals that are completely mismatched than two that are completely matched on their non-introducing dimensions. The greater effort is needed because there is more information and a harder attribute binding problem for a mismatched dimension where the fact that there is a mismatch must be remembered as well as which individual the different properties belong to. We claim that this extra effort is taken up in making more and more complex associations. Errors in recall are caused by the disruption of one or more features. A candidate model for the way in which a disrupted representation might cause a recall error is discussed in Chapter 5 where a PDP network is constructed that is capable of correct recalls but makes errors when subjected to a corrupted input representation.

The model contrasts with most of those described in Chapter 1 by not treating property attribution as an unanalysable primitive. Rather than simply linking a unitary individual and a unitary property, binding produces a fragmented and redundant representation that requires an inference to be made to reconstruct it. The redundancy of the representation contrasts with the non-overlapping fragments of Jones' work. Jones' model only deals with single individuals whilst our model copes with pairs of individuals and could be extended to further individuals.

In the text modes used in the experiments described in this thesis, the order of descriptions of the pair of individuals is clear and predictable — the descriptions are either blocked (IxI) or alternate (PxP). Presumably, subjects are able to take advantage of these temporal cues when they make the associations necessary for correct recall. If the temporal cues were to be disrupted the model would predict that greater mnemonic effort would be needed and we would expect a more complex array of features to be picked out in the regression model. Stenning, Patel and Levy (1987) describe a task where some of the text modes are very temporally unpredictable. Their recall task asked subjects to recall the individuals in a fixed order — either first individual introduced followed by second individual or vice versa. [1] They divide their material into two groups. In the first group the cueing instructions were obeyed. In the second, the text modes were more complex and subjects had great difficulty obeying the order of recall instruction. It seems that subjects become confused about which individual was introduced first for these more complex modes and thus are unlikely to be able to use temporal cues very well. The regression models support this assertion since the model for the second group is more complex, containing a greater number of features.

Another prediction of the recall model is that the associations made are likely to be affected by how stereotypical the descriptions are. This idea is currently being investigated by our research group.

3.13 Implications of the final model

The instantiation model has more and larger intra-individual features than the tagged model and accounts for the data as well as the tagged model. The main difference is a theoretical one — the instantiation model requires an extra mechanism to fully achieve binding, while the tagged model begged the question by simply labelling each intra-individual feature with the individual it belonged to.

The claims of the instantiation model are that properties are bound to individuals

[1] Naturally, this recall order was varied randomly.

indirectly through inference from a redundant fragmented 'database' of existential 'facts'. Each 'fact' is an association recruited from already existing knowledge that provides mnemonic support for the representation of the information required for the task in hand. For the MIT, this task is usually restricted to one of remembering the correct assignments of properties to individuals, rather than the actual properties themselves which are given by the menu.

Because the database is redundant, if part of it is corrupted it will tend to become inconsistent (see Figure 3.4). We can suppose that this process is what leads to errors in recall. Potentially, the system should be able to correct some disruption if the inconsistency can be resolved. It would be instructive to be able to investigate a model of these inconsistency resolution and error correction processes.

At this point it is worth comparing the model to Jones' Fragmentation Model (see Chapter 1). The material he uses only ever describes a single individual and so there is no problem for the subject concerning which individual to attribute a particular property to for a particular trial. However, since memory is tested after all the trials have been presented, the subject does have to contend with possible interference between trials. This interference is minimised by the fact that none of the descriptions overlap. These details of the task contrast with the overlapping descriptions of pairs of individuals presented in the MIT.

Jones' methodology is to probe memory by cueing with a varying number of the original attributes of a trial (Colour (C), Location (L), Object (O) and Sequential position (S)). From the frequencies of correct cueings, Jones infers the frequencies of different fragments in memory. His fragmentary representations are somewhat similar to the features described here except for the fact that they are totally non-redundant, i.e within a single representation, fragments do not overlap. It may be that subjects do not need to employ redundancy for such simple descriptions or perhaps his modelling methods are too conservative to pick up extra overlapping fragments.

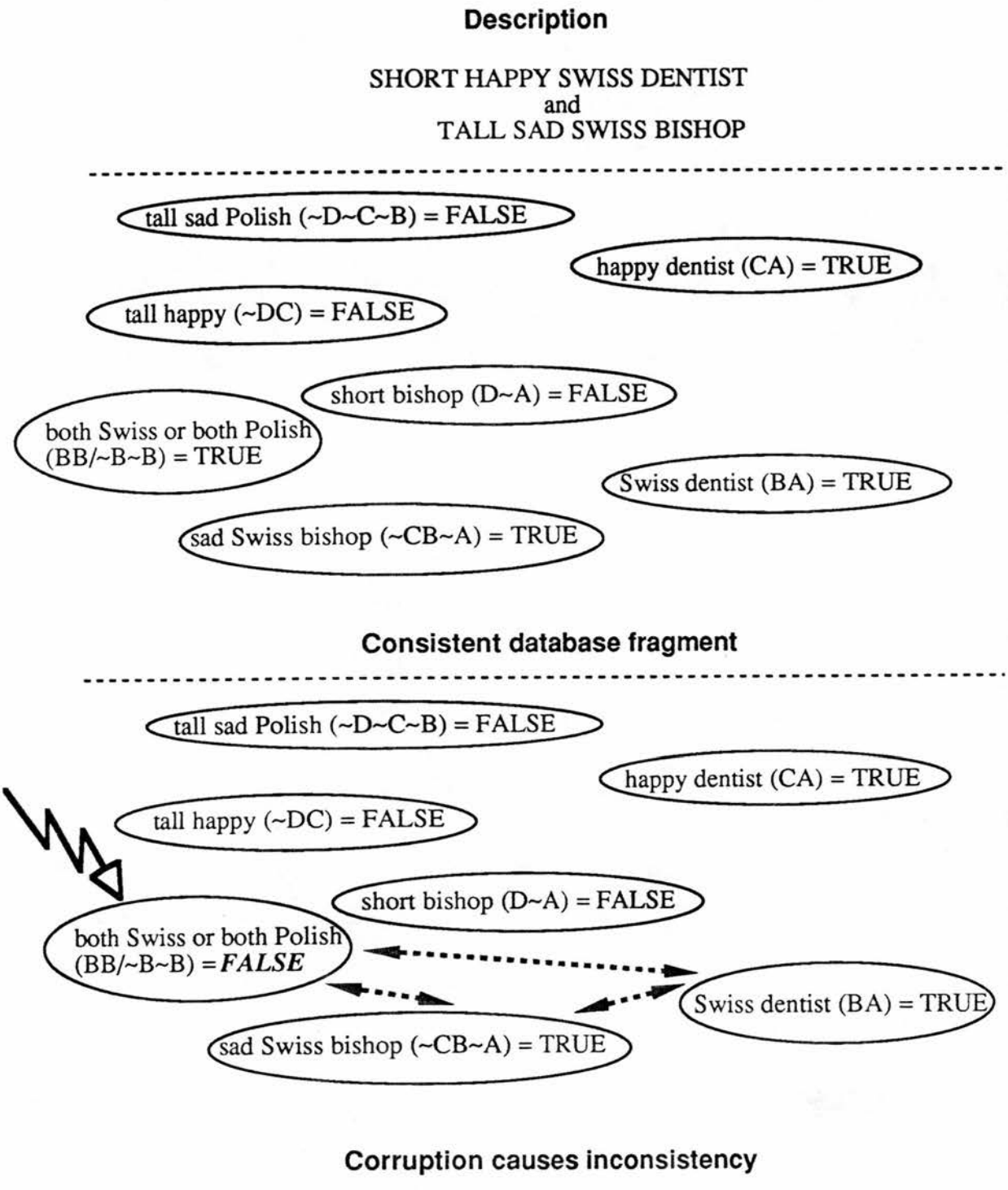


Figure 3.4: Inconsistency due to corruption of the Representation

3.14 Paving the way to a PDP model

As its stands, the statistical procedure has extracted a similarity metric by choosing a set of features, assigning them coefficients and using them to account for the frequencies of certain classes of error. The value of the technique has been to specify a redundant fragmented representation. However, the model does not explain how a system might do the inference from fragmented representation to correct recall or how such a system could produce a well-formed recall (whether correct or in error) when the underlying representation had been made inconsistent by some disruption.

What is needed is a process model based on the feature set extracted by the statistics that is capable of explaining how a system can perform the above processes. By this, we mean a process model in the weak sense of a cognitively plausible computational mechanism capable of performing this function. The PDP networks described in Chapter 5 achieve this. There are other benefits gained from extending the statistical model using the PDP framework. These general attractions of PDP are discussed in Chapter 4.

Chapter 4

Parallel Distributed Processing

4.1 Introduction

This chapter gives a brief introduction to the aspects of the Parallel Distributed Processing framework that are relevant to the work described in the thesis. The attractions of the general framework are discussed and the background details necessary to understand the network model in Chapter 5 are described. The treatment given here will be more of a concise justification of the use of a PDP modelling framework than a detailed introduction to the field. Relevant general introductions to the field can be found in Rumelhart and McClelland (1986), Johnson-Laird (1988) and Levy (1988). A varied collection of important papers can be found in Anderson and Rosenfeld (1988).

4.2 What is Parallel Distributed Processing?

I shall use the terms Parallel Distributed Processing (PDP), neural networks and connectionism interchangeably. Distinctions can be made but the fundamental concepts behind all these terms are the same. The differences usually reflect the way in which the modelling framework is applied.

There has been a recent resurgence in interest in the abilities of networks of

very simple computational units, connected in such a way that the network as a whole performs some useful processing task. The roots of the field can be traced back to James (1892), McCulloch and Pitts (1943), Rosenblatt (1958) and others. The recent research effort is due to the energetic and well publicised research of such people as Rumelhart and McClelland (1986), Hinton and Anderson (1981), Sejnowski and Rosenberg (1986), and Hopfield (1982).

We will now informally describe the makeup of a neural, connectionist or PDP network. Typically, these networks consist of a number of units and weighted connections. Each unit performs a simple summing operation on its inputs and outputs a function of this summed input. The most usual output functions are linear, threshold or sigmoid (see Figure 4.1). The links or *connections* between the units carry a positive or negative weight. The input to each unit from a particular connection is usually equal to the output of the 'sender' unit multiplied by the weight that has been assigned to the connection (see Figure 4.2). There can be any pattern of connections between the units. A common architecture is one where every unit is connected to every other unit. Another is one where there are layers of units, each connected only to the next layer (see Section 4.5).

The computational mechanism in these networks is the dynamics of the flow of activation through units and connections. The networks can perform a computation if the states of single units or a pattern of units can be interpreted as representing something and the dynamics of the network cause it to evolve from one pattern of activation to another. Often, the network as a whole will have a number of *stable states*. These are states of the network that tend to remain the same and to which the network tends to evolve if it is close to them in *state space*. State space is the abstract space with as many dimensions as there are units in a network. A network's position in state space is defined by the activation levels of each of its units. The process of evolution from an unstable point in state space to one that constitutes a stable state is often termed *relaxation*.

A paradigm case of relaxation is one where a stable state can be said to represent a stored 'memory'. If a pattern of activation representing a unique fragment of

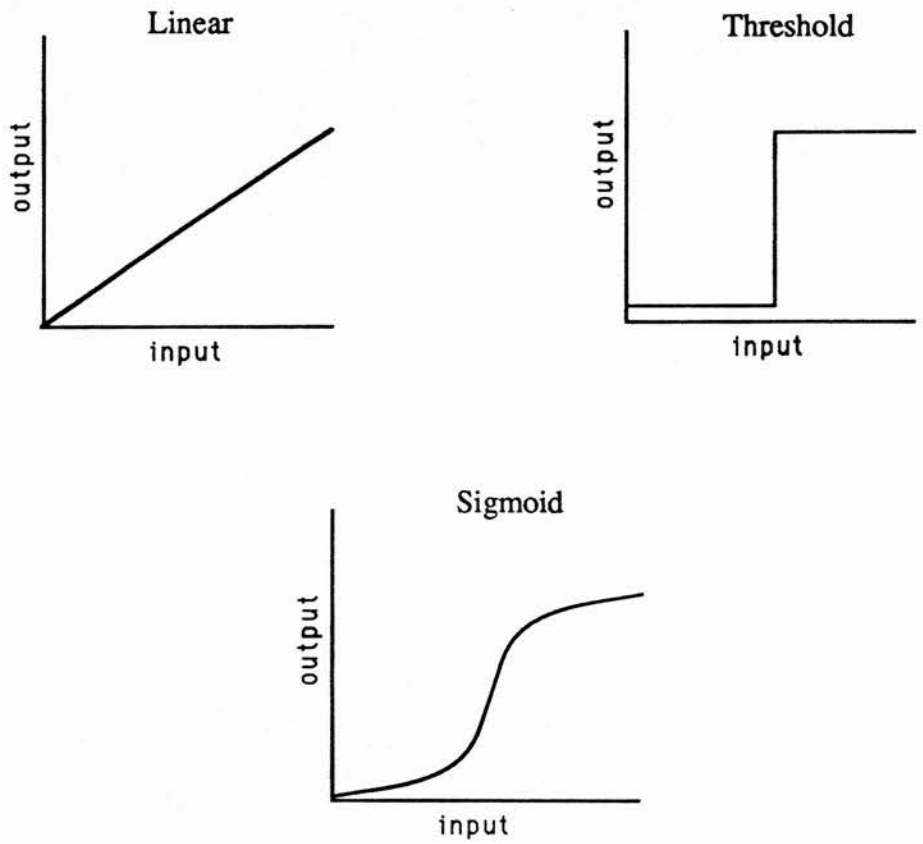
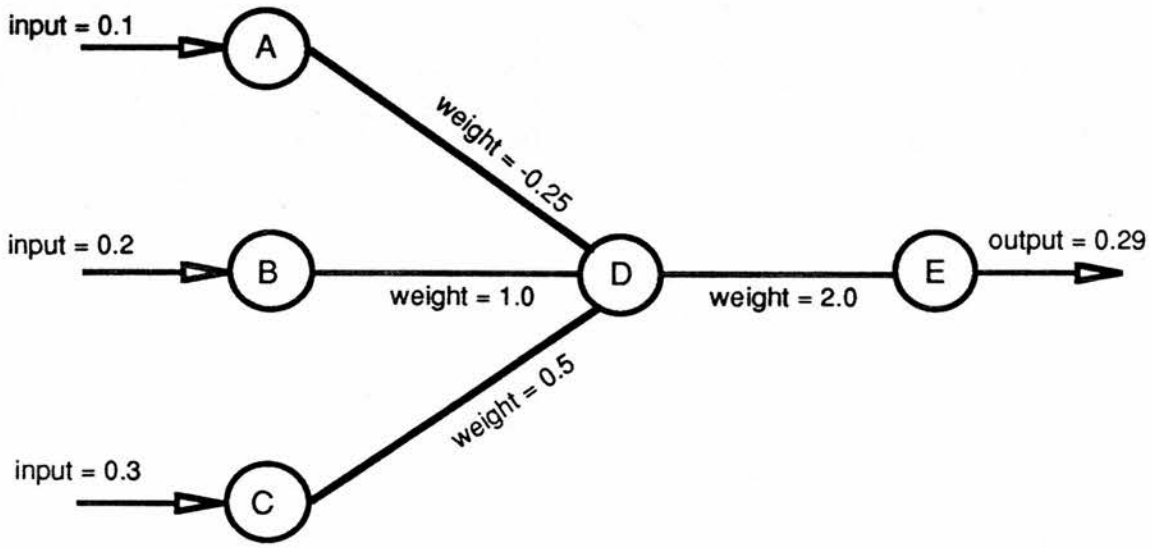


Figure 4.1: Some examples of output functions



If each unit outputs the sum of its inputs then:
The input to unit D = (0.1 x -0.25) + (0.2 x 1.0) + (0.3 x 0.5) = 0.145
and the output of unit E = 0.145 x 2.0 = 0.29

Figure 4.2: A simple network

the pattern of activation that the network has in the stable state is imposed on the network, it will relax into the stable state. This process can be viewed as one of pattern completion, content addressability, cued recall or noise resistance, depending on the nature of the initial state of the network.

In mathematical terms, for a network to behave in this way, its overall state can be described in terms of an 'energy' and its dynamic behaviour is such that the energy is minimised. This can be visualised in simplistic terms by thinking of the state of the network as a ball bearing on a slope whose behaviour is such that it minimises gravitational energy by rolling down the slope. Some important results for the energy functions of simple networks have been proved by John Hopfield for simple networks whose nodes have discrete (Hopfield 1982) or continuous (Hopfield 1984) values.

The single stable state of a system described above for a network is known as a *point attractor* in the theory of dynamic systems. *Cyclic attractors* are sequences of meta-stable states. There is a lot of interest amongst connectionist researchers in ways of making networks go through sequences of states (e.g. Pineda 1987, Elman 1988) since many possible applications involve the processing of temporal information e.g. robotics and speech processing.

The dynamics of a network are controlled by the values of the weights of its connections. It is these that control how much 'excitation' or 'inhibition' is passed round the network. The values of the weights effectively shape the 'energy landscape' through which the state of the network travels. PDP networks usually have several stable states, corresponding perhaps to different 'memories'. The storage of these memories is *distributed* across many weights and the same weight can take part in the storage of many memories. Depending on what the network is being used for, the values of the weights can also be viewed as controlling the way information is processed since they control how one pattern of activation in the network is transformed into another. In this context a single weight can be said to act as a 'constraint' and the process of relaxation to be one of 'constraint satisfaction'. Since many weights are used to store the information in a network, there

can often be a built-in resistance to damage and noise — performance may not be perfect but will not break down catastrophically (i.e there is 'graceful degradation'). In ideal cases, noise or slight errors can be completely eliminated by the network relaxing into the 'correct' stable state from a noisy starting state.

One of the most attractive properties of these systems is their ability to learn. Learning amounts to a method of adjusting the weights so that the network has the correct stable state for a particular memory or the correct output for a particular input. The obvious difficulty is that the weights must be adjusted so that *all* the memories or *all* the input-output pairs are learned. There is an upper bound in the number of items that can be stored in a given set of weights before performance breaks down and items are confused. This 'blurring' when a network is overloaded is not necessarily an altogether unattractive property (see Section 4.3). Details of various training algorithms can be found in Rumelhart and McClelland (1986).

4.3 The Attractions of a PDP Approach

There are many processing and representational properties of PDP networks that make them attractive as a framework for the modelling of cognition. These properties emerge naturally from the way that PDP systems work. They can be thought of as new modelling primitives or as new metaphors for the discussion of cognitive processes (Norman 1986).

The natural way in which content addressability arises from the structure of simple PDP networks makes the framework an attractive one for the modelling of human memory. The same computational mechanism of relaxation into a stable state from a 'nearby' state can be used as a model for cued recall, associative memory and resistance to noisy data.

The conception of a weight as a constraint on the trajectory of the network through state space is a very powerful one for the cognitive modeller. It can be used to model the use of background knowledge in memory or the use of context in

language understanding. 'Knowledge' stored in this way, whether it is a particular past experience or the way to transform a given type of pattern to another one, is based on 'soft' or 'weak' constraints (Blake 1983) rather than the idea of the explicit rule taken from logic, serial computation and natural language syntax that has been so dominant in cognitive psychology, cognitive science and artificial intelligence.

Another attractive property arises from the blurring process alluded to above. If certain types of networks are overloaded or fed patterns with too great a degree of overlap, they store a composite or generalisation of the patterns. This has certain attractions to those who wish to model the learning and representation of conceptual knowledge (see Section 4.5).

In contrast to serial architectures, PDP models seem to exhibit similar weaknesses to those of human cognition. Both can be shown to have limited immediate storage capacities if they are presented with several stimuli that are strongly similar along the dimensions that they are being coded with (Miller 1956, Conrad 1964). The use of unbounded (or indeed relatively shallow) recursion in the processing of linguistic structure poses problems to humans and does not appear to be a natural mode of operation for PDP systems (Miller 1962, McClelland and Kawamoto 1986). Importantly, these weaknesses are not arbitrary limitations added to explain human fallibility but natural characteristics of PDP networks.

4.4 Objections and Problems

The PDP framework, despite its age (see James 1892, Anderson and Rosenfeld 1988), is proving to be a radical departure from previous modelling frameworks for cognitive psychology. It is probably true to say that it has not yet proved itself but has shown considerable promise. It has certainly provoked controversy and this section will mention some of the principal points of contention. The section will end with a brief discussion of some of the technical problems that must be confronted and overcome if PDP is to be a success.

One of the most reported and argued over articles to appear in Rumelhart and McClelland (1986), one of the widest used references in the field, was the chapter by Rumelhart and McClelland called 'On Learning the Past Tenses of English Verbs'. They presented a simple model that appears to go through the same major stages in the learning of regular and irregular past tense formation that children go through — namely, the learning of irregular forms such as 'came' and 'went' followed by the learning of the general rule to add the -ed suffix which is initially applied to the irregular forms as well (e.g 'camed'). It was claimed that a PDP approach of capturing this phenomenon using constraints learned from examples worked as well as one based on the formulation of explicit rules. Pinker and Prince (1988) have given a detailed attack on the model. They show that it is deficient in accounting for some of the psychological data and could not in principle do so in its present form. They attack the whole PDP framework for being, as yet, incapable of accounting for the sort of behaviour that is currently modelled using explicit rules. As Smolensky (1987) argues, it is perhaps slightly premature to dismiss the whole connectionist approach on the basis of the deficiencies of an early model. The real issue at stake is not whether PDP systems are capable of emulating systems of explicit rules, but whether models employing the natural properties of PDP systems are going to be useful. Rumelhart and McClelland's claim that apparently rule-based behaviour emerges from a conspiracy of soft constraints is one that deserves to be taken seriously (see Smolensky 1988, Johnson-Laird 1987).

There have been various other criticisms of the PDP framework that might be loosely defined as attacks on its lack of representational sophistication. Norman (1986) worries about the difficulties that PDP has in representing the type-token relation and variables. Fodor and Pylyshyn (1988) consider that PDP can only in principle be a basis for the implementation of cognitive models but not for theories of 'cognitive architecture' because its representations lack internal combinatorial structure. These issues are not being ignored (see Touretsky and Derthick 1987, Hinton 1987, Smolensky 1988). Perhaps they represent the beginnings of a Kuhnian paradigm shift away from explicit rules and symbol processing and towards a new science of what David Rumelhart has called 'brain-style' computation.

Although there is obviously a much stronger similarity between PDP architectures and neuroscience than there is for traditional serial computer architectures, cognitive modellers should beware the distraction of biology. The interdisciplinary nature of cognitive science strengthens it but the admission of neuroscience to the club should not be allowed to detract from a proper psychological level of description for cognition. It is a step forward that we can now, to some extent, communicate with a common vocabulary and draw on some of the same metaphors but we are not yet ready to integrate physical, chemical, biological and psychological levels of description. In other words, we are not yet at a stage where a cognitive model can be dismissed on the basis of its lack of neuroanatomical or neuropharmacological realism.

There are a number of practical problems concerning the use of PDP as a modelling framework. Most models in the literature solve a limited problem by means of a dedicated network module. There is a good understanding of how traditional serial modules can communicate and be controlled, but there is a lack of such an understanding for PDP systems. Mental processes will not be understood by a single large network model. We need a better understanding of how to model the flow of information and control between different modules (see Norman 1986).

There is a practical problem in the modelling of large networks. Current training methods such as back-propagation or, worse still, the Boltzmann machine (Hinton and Sejnowski 1986) are slow and do not scale well. Until better methods are developed we have to rely on developments in computer technology such as faster processing units and parallel architectures if we want our simulations to take a reasonable amount of time. Alternatively, we can use hardware implementations of connectionist networks (see Sivilotti et al. 1987, Murray, Tarassenko and Hamilton 1988). Building physical neural networks is hard because of the number of programmable connections that are needed. An alternative to VLSI is the use of optical techniques (e.g. Farhat et al. 1985), where the use of light beams to implement connections may solve the wiring density problem.

4.5 Two Representative Types of Network

To make the discussion somewhat more concrete we will now describe two representative network architectures. The first, a simple auto-associator, can demonstrate many of the attractive computational properties described above. The second is computationally more powerful and is used for the models described in the following two chapters.

An excellent example of a model based on a simple auto-associator is the model of human learning and memory reported in McClelland and Rumelhart (1985, 1986). The network used was one where every unit is connected to every other unit but not to itself. [1] The units take continuous activation values ranging from -1 to +1. Each unit can be turned on by an external stimulus as well as being influenced by its fellow units. The weights of the connections can take any real positive or negative value. The activation function used is a simple sigmoid augmented with a decay term that tends to push activation towards zero. The training rule used is the so called *delta rule* that changes the weight to a unit by an amount proportional to the difference between the actual output of the unit and its desired output.

It is easy to show that this type of network can display content addressable properties. Rumelhart and McClelland were concerned in modelling the coexistence of general and specific knowledge. They invented an arbitrary pattern to represent a "dog". The network was trained on random distortions of the 'prototype' that represented specific exemplars. The stable state of the network after this training was close to the prototype — the network had generalised from the exemplars and recovered the prototype. The same network was also capable of recovering three different prototypes for "dog", "cat" and "bagel" patterns. The dog and cat patterns were not orthogonal since they represented somewhat similar categories. If a few exemplars were repeated relatively often the network was able to store

[1] Connecting a unit to itself tends to be destructive because during training the weight on this connection tends to get very large since it is easy for this weight to learn to maintain the activity of the unit. What is required during learning is that the activity of the other units maintains desired level of activation in a particular unit.

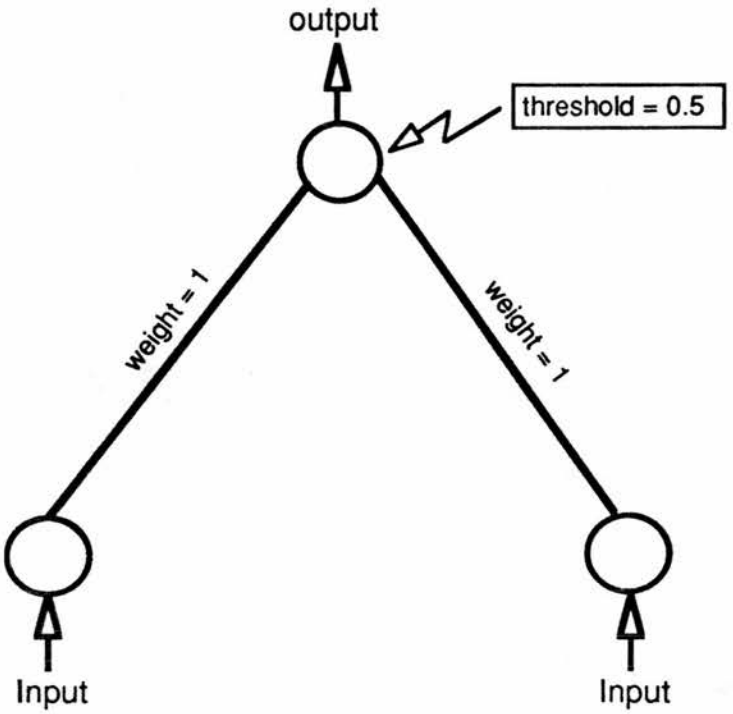


Figure 4.3: A network capable of computing the logical OR function

a prototype and two specific exemplars — a generalised dog as well as Fido and Rover.

The authors went on to simulate several other phenomena from the literature using the same network. They also discuss augmenting the model with extra 'hidden' units. We will now describe a network architecture where such units have proved to be very useful.

Layered networks can be used for several purposes. Usually, there is a layer of input units and a layer of output units. The networks can be trained to associate a given output with a given input. Sometimes this amounts to the network learning a simple mathematical or logical function. An example might be a network that learns to perform a logical OR function (see Figure 4.3). Another task trained in the same way might be one of classifying inputs into various different categories. It has been shown (Minsky and Papert 1969, 1988) that a procedure exists that can train any such two-layer network to perform any task that it is capable of. They also demonstrated that this type of network was not capable of learning the

exclusive-or function (A or B but not both) without an extra layer of hidden units that are neither used for input nor output. Many more interesting problems can be shown to be equivalent in difficulty to the exclusive-or function. At the time of writing there was no known training procedure for this sort of network and so Minsky and Papert were widely blamed for the loss of interest in connectionism during the 1970s.

Recently, a training algorithm for multi-layer networks has become widely known and used (Rumelhart, Hinton and Williams 1986). Similar methods have been described by Werbos (1974), Parker (1985) and Le Cun (1985). The method used is called back-propagation because it allows the delta rule to be generalised for non input/output layers by passing back an error term from the output layer to the input layer via the hidden layers so that intermediate weights can be altered. This algorithm works for layered networks where activation can only flow forwards towards the next layer. There is no feedback from one layer to a previous one and there are no connections within a layer. Connections are allowed to 'miss' a layer or layers.

The use of this training method has allowed the use of more powerful network models (e.g. Sejnowski and Rosenberg 1986) and has promoted interest in connectionism in general. Without the existence of this training method the network models described in this thesis would not have been possible.

4.6 Conclusions

Parallel Distributed Processing gives us a variety of useful primitives for the modelling of cognition. It is a young field and is not without theoretical and practical problems but does show considerable promise. The PDP modelling framework was chosen for the research described here because of its general attractions. The layered network architecture used in the following chapters does its job well but it is just as important that the model can be related and potentially integrated with models of other aspects of cognition within a general framework.

Chapter 5

PDP Models of Recall Processes

5.1 Introduction

This chapter describes how the linear statistical model of error data was used as the basis for the development of PDP models of recall performance. The networks were designed to show how a simple system could synthesise the information from the fragmented underlying representation into a recall of a pair of individuals. The models were then able to show how errors resulted from such a system when the feature representation was disrupted. Since the feature representation involved considerable redundancy, any such disruption is likely to cause inconsistency. To make a well-formed response, whether correct or not, the network must be able to resolve any inconsistency.

The reasons for extending the statistical model will be examined and the theoretical insight that was gained will be discussed. The way in which the modelling enterprise developed will be demonstrated by the description of the two main models. The first model was trained to output one of the two correct orders of individuals from consistent input vectors. The second network was trained to produce both of the possible recall individual orders from the same input vector with the addition of a 'cueing unit'.

The chapter will describe and justify the decisions made in the choice of net-

work architecture, training regime and simulation trials. The problems that arose in the use of the standard training algorithm and the way in which they were circumvented will be outlined. The particular way in which the networks were tested for their generalisation abilities will be described. The general aspects of the modelling framework will be described first and then detailed descriptions of the development of the two network models will be given. The chapter ends with a general evaluation of the usefulness of this modelling approach.

5.2 The Modelling Framework

The linear model extracted by multiple regression and described in Chapter 3 accounted for most of the variance in the log adjusted frequencies of the recall error categories. The model consisted of independent and redundant structural features whose truth values specified the properties of a pair of individuals. What was lacking in the statistical model was a performance mechanism by which a recall could be achieved from the fragmented representation specified by the statistical model. This calls for a process model that can make the correct inference from the truth or falsity of the existential facts in the 'database' to the correct specification of the eight properties of the pair of individuals in the original description. Once such a model is constructed, the possible mechanisms that cause errors can be examined. Since the database is redundant, disruption is likely to make it inconsistent, i.e., the facts will not be consistent with any possible pair of individuals. The inferential mechanism must be able to cope with the introduction of inconsistency. It may not be possible to correct the error since there may be several equally likely well-formed vectors. However, even if it is impossible to correct the error with any certainty or if the mechanism is not capable of a possible error correction, it must be able to resolve the inconsistency and produce a well-formed recall, just as a subject has to do when faced with a recall menu.

It was felt likely (and was empirically verified) that the inferential task of mapping feature truth values to property specifications would require the computational power provided by at least one hidden layer of a PDP network. The exact math-

ematical proof of this requirement is complicated by the redundant nature of the input, but tests with two layer networks demonstrated that they were not capable of the learning task for even the first network. A three layered feedforward network trained using 'back-propagation' (Rumelhart et al. 1986) was chosen since this architecture and learning rule is capable of learning the correct weight values in such a network in a reasonable time span. A 'Boltzmann Machine' (e.g. Hinton and Sejnowski 1986) would be capable of learning the same task but the learning algorithm is extremely computationally intensive and takes a very long time in practice. Feedforward networks by definition do not allow any feedback and so are unable to 'relax' into the nearest stable state. This means that they do not exhibit the same pattern completion properties as an auto-associator (see Chapter 4). The inferences required in the models described here can be carried out without pattern completion. The networks are models of the production of retrievals from a given database rather than models of the organisation and behaviour of the database. The weights in the networks described in this chapter store the constraints that conspire to perform correct inferences and maintain consistency; they don't store the underlying representation. The representational scheme has been already specified by the statistical model. Rather than modelling memory storage processes, the networks described here model one aspect of how fragmented and redundant information is processed during recall.

The input layer of the network was used to represent the truth values of the individual underlying features. The output layer represented the specification of the eight properties of the pair of individuals. Such a PDP network can be viewed as a constraint satisfaction system and it was hoped that it would be able to produce well-formed outputs even when its inputs were disrupted in a way that made them inconsistent. Although the use of general knowledge is not explicitly represented in the models described in this chapter, the use of a PDP framework may eventually allow the modelling of content effects as extra constraints or connections. The networks described here represent the existence or absence of a combination of vocabulary items, and hence the truth or falsity of a particular feature, by the binary activation level of a single node. A more ambitious model might be based on a more distributed, less abstract representation that included constraints from

general knowledge. Recent work in our research group has been aimed at characterising some aspects of subjects' knowledge of the vocabulary used in these experiments in the form of simple PDP networks (e.g. Nelson 1988).

5.3 The Nuts and Bolts of the Simulations

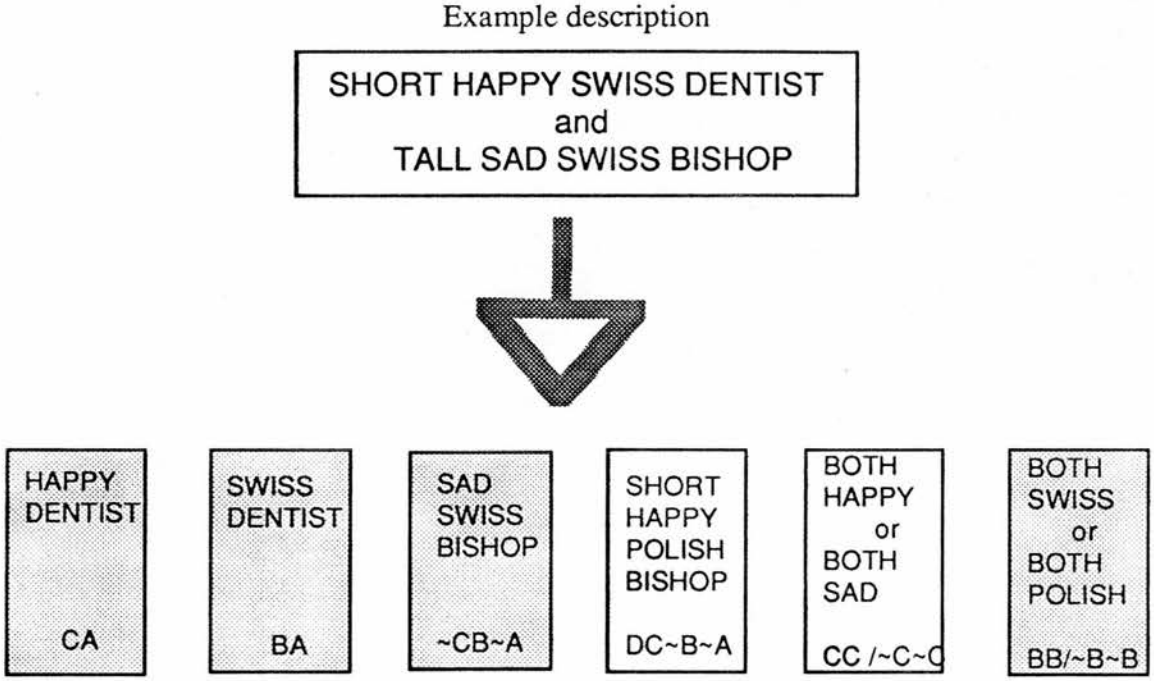
Before describing the two simulations in detail, the basic modelling approach will be outlined. This includes the architecture of the networks used, the way they were trained on well formed stimuli and the way errors were induced by injecting noise into the input layer. The practical problems of choosing appropriate training coefficients and avoiding 'stuck' output units will be discussed in some detail. The way the representations used by the networks were tested for generalisation abilities is also introduced.

5.3.1 The network architecture

The architecture of the networks is essentially the same as the one described in Rumelhart et al. (1986). The units have a continuous sigmoidal activation function based on the logistic function. Each has a bias term that can be viewed as a weight from a unit that always has an activation of 1.0. The layers are strictly feedforward and connections are not allowed to miss layers.

The networks consisted of three layers. The activation of the units in the input layer represented the truth values of the features extracted by the regression model. Each feature was represented by the activation of a single unit. These units, then, only took values of 0 (false) or 1 (true) (see Figure 5.1).

The activations of the eight units in the output layer (see Figure 5.2) represented the values of the eight properties of the pair of individuals for a particular menu. Again, these take values of 0 or 1 for a well formed case, so a 'bishop' might be represented as an activation of 0 and a 'dentist' as a 1, and 'fat' as 0 while 'thin' is represented by 1. It might be expected however that when the input layer vector



Sample of units in the input layer

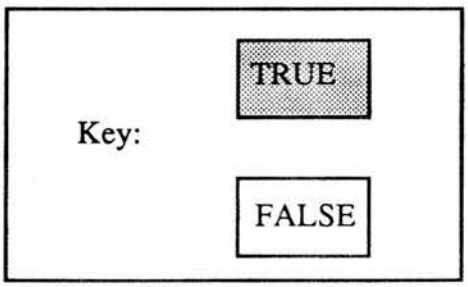
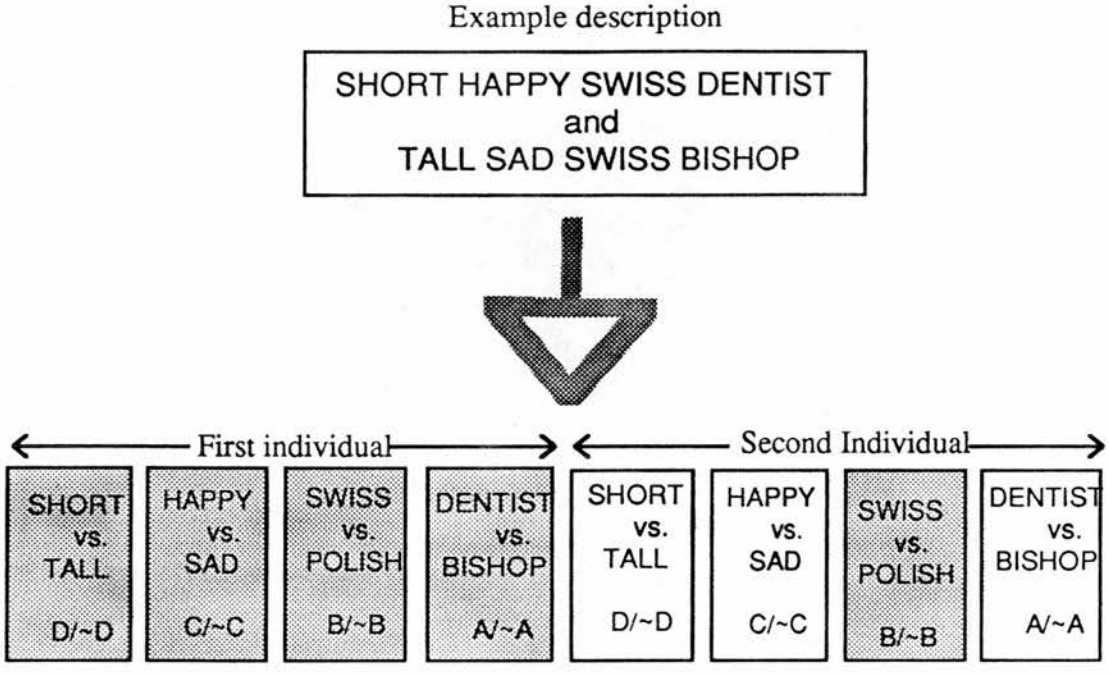


Figure 5.1: The input layer



The activation of the units in the output layer

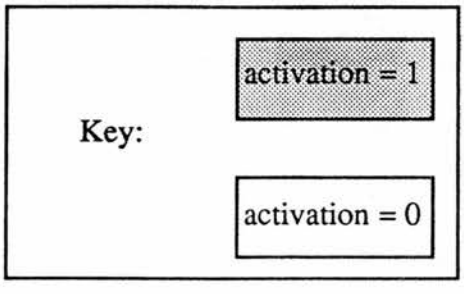


Figure 5.2: The output layer

is disrupted the values of the activations in the output layer would depart from their strictly binary nature since they can take continuous values between 0 and 1 and this sort of network is not able to settle into the nearest stable space in activation space. The extent to which this happens is discussed in the detailed descriptions below.

The number of hidden units required is hard to judge. Within reasonable bounds, the smaller the number of hidden units the longer it takes the network to learn. Too few hidden units can result in a network whose learning curve fails to converge. However, generally speaking, a constrained number of hidden units leads to better generalisation properties (see Wieland and Leighton 1987). The number of hidden units for the networks below were chosen as a compromise between these two factors using experimentation and the experience of many simulations as a guide.

For generality and simplicity, the networks were fully connected i.e. each unit in the input layer is connected to each one in the hidden layer, and each unit of the hidden layer is connected to each one in the output layer.

It is clear that the specification of the input layer of the network is directly specified by the statistical analysis of psychological data. It is rare for the specification of a PDP model to be defined by real data. It is particularly interesting that the distributed and redundant nature of the representation extracted by the multiple regression procedure should suit a PDP framework so well. In some respects this is no surprise since the underlying matrix algebra of multiple regression and linear networks, at least, is very similar. The models described here are novel hybrids. The input layer representation is extracted from psychological data by a statistical procedure. The output layer representation is specified by the programmer to be the simplest way to encode the descriptions used in the experiment in terms of the properties given in a particular menu. The network's task is to encode a representation in its weights that can map the input representation to the output representation. Usually, both input and output representations are specified by the programmer. An interesting exception is the room schema model described by Rumelhart, Smolensky, McClelland and Hinton (1986) in which the weights

themselves are specified by a statistical manipulation of subjects' judgements. It is important that the PDP modelling paradigm should be able to accommodate psychological data, rather than be used purely to demonstrate potentially interesting computational and representational properties.

The two network models described in this chapter are the simplest possible extensions of the statistical model. The values of the regression coefficients and mean accuracy of the features in the human data are not used to define the network architecture or to influence the way that the input layer representation is disrupted by noise during simulation runs. The network is based only on *which* features were picked by the multiple regression procedure. It would not be surprising if the network model does not fit the data exactly. Since there are so many parameters that can be manipulated in a PDP model, it is important to start with as simple a model as possible and investigate its behaviour fully, before adding extra parameters. The purpose of the PDP modelling is to investigate plausible representational mechanisms rather than to provide a mere description of the data by accounting for its variance. The statistical account of the data provides an important foundation for this modelling process by providing a starting point that does account for most of the variance in the data.

5.3.2 The training regime

The network was trained using a slightly modified version of the standard back-propagation learning rule (Rumelhart, Hinton and Williams 1986). Each member of the training set is presented to the network and the error term and weight increment for each connection is calculated. At the end of every sweep through the pattern set (epoch) the weights are adjusted. The training is terminated when none of the output units for any of the patterns in the training set have an error of greater than 0.2. This value was chosen as a compromise between perfect learning and a reasonable training time. This condition defines what will now be referred to as *reaching criterion.*

It was decided to train the network on all the appropriate input-output pairs whose

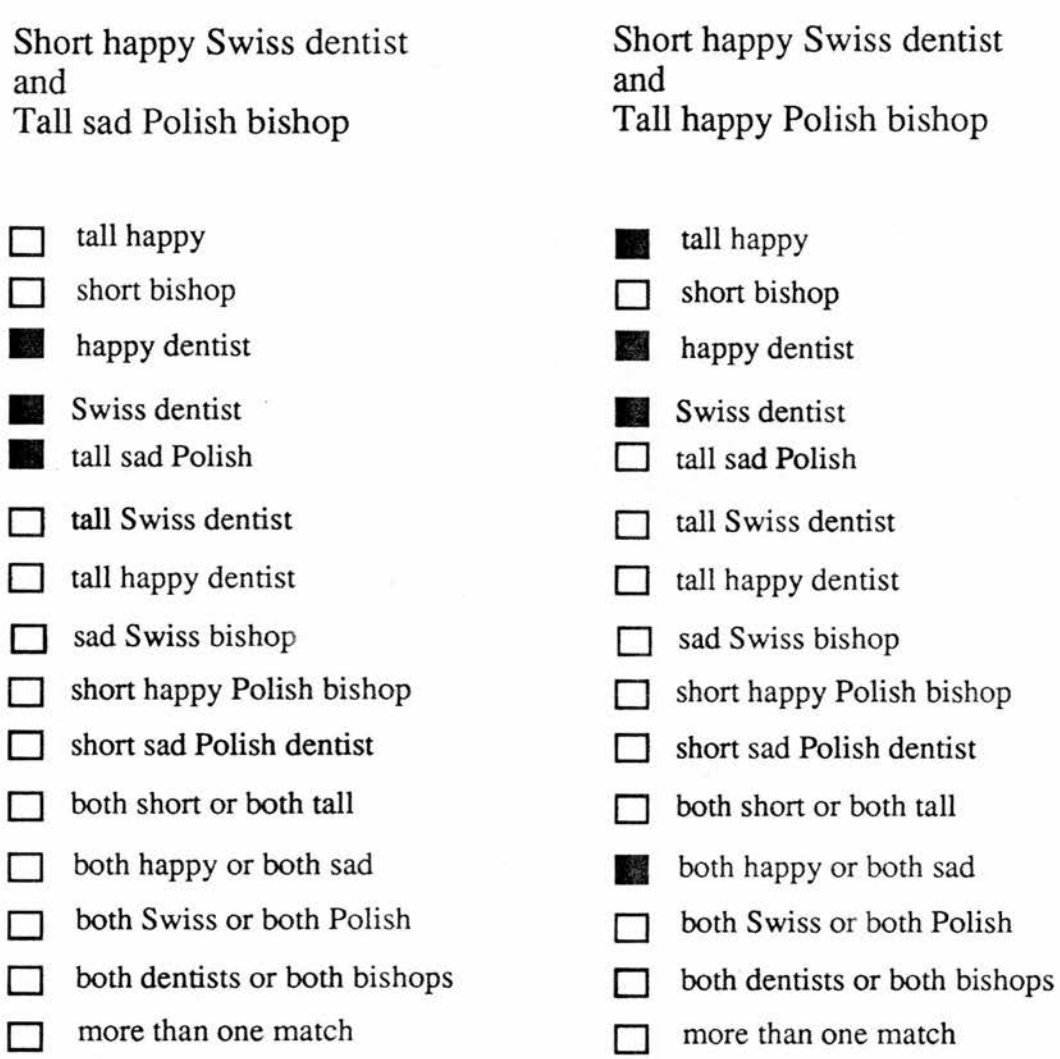


Figure 5.3: Two descriptions that are distinguished by their feature value vectors

inputs were distinguished by the features of the recall model (see Figure 5.3). In other words, all distinct mappings between feature value vectors to property values were used including those for descriptions with matching introducers (e.g. a short sad Polish bishop and a tall happy Swiss bishop) and those for pairs of identical individuals (e.g. a short sad Polish bishop and a short sad Polish bishop). There were three pairs of training patterns where both feature vectors were identical and thus not distinct (see Figure 5.4). All six of these patterns were ones describing pairs of identical individuals [1]. One of each pair was arbitrarily chosen

[1]The recall model described in Chapter 3 was logically complete for the stimuli used in the

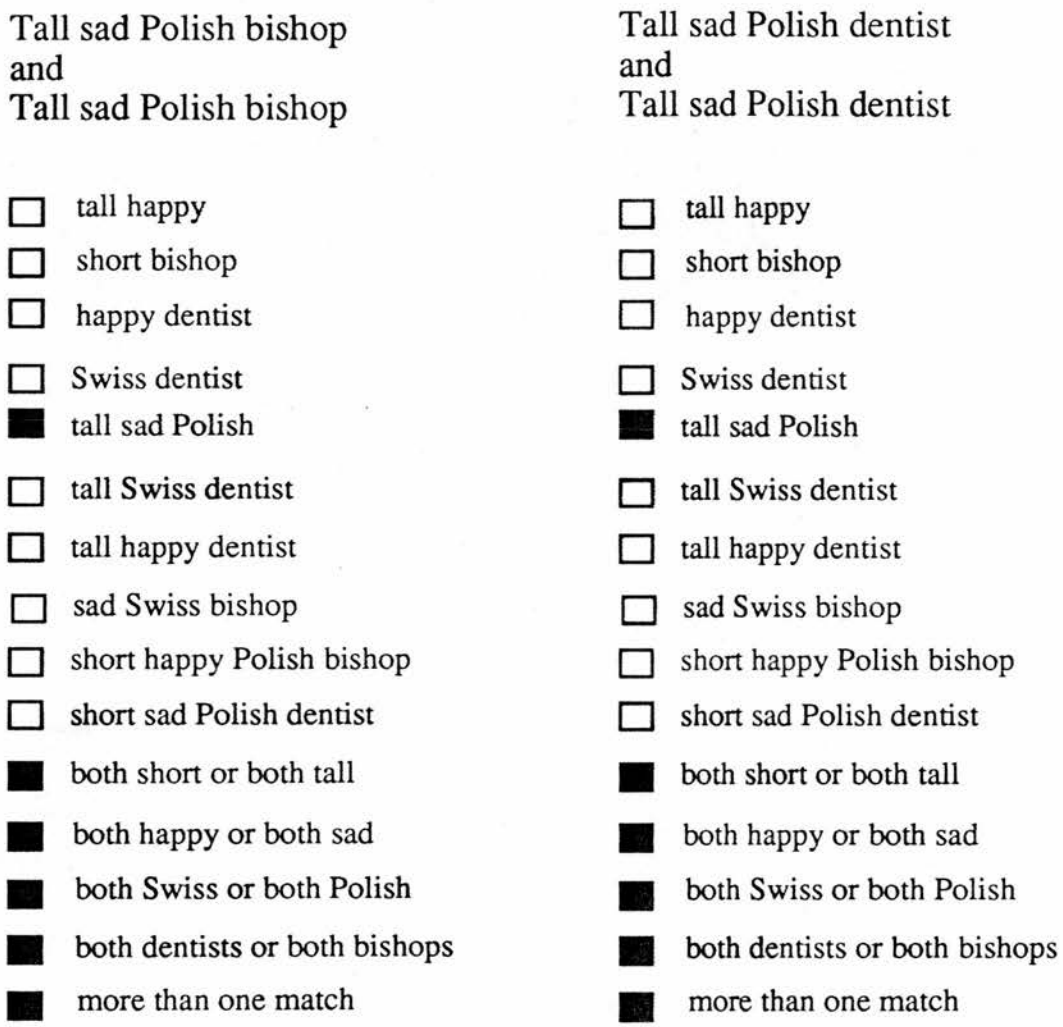


Figure 5.4: Two descriptions that are not distinguished by their feature value vectors

to be discarded since it is impossible to train a network of this kind to produce more than one output (property value vector) from the same input (feature value vector). This left 253 possible items for training sets. 120 of these corresponded to the stimuli that were used in the experiments, i.e. where the introducing dimension was mismatched. A further 120 described pairs of individuals with matched introducers. The remaining 13 described completely identical pairs of individuals. These latter 133 training items could have been left out of the training set if the nets were to be trained only on patterns that described well-formed stimuli from the experiment. However, pairs with matched introducers and identical pairs could be made as responses by the subjects (though this was rare) and so they were included in the training set to allow this type of response even if it is always incorrect. It was also felt desirable to teach the networks the complete 'logic' of the relations between properties, including matched introducers, even if they never occurred in the experimental stimuli. Subjects were presumably able to envisage such descriptions.

Parameters used

Apart from the architecture of the network and nature of the training set, the training of this type of network is very dependent on the values of its learning rate, η, and momentum, α, coefficients. If these are set wrongly the learning can oscillate wildly or even 'lock up' and stay at a high error rate for ever. The actual values of these coefficients were obtained by trial and error and the experience of running many simulations.

Particularly for networks whose computational power is restricted by a small number of hidden units, finding appropriate values for the coefficients can be very time consuming. Even if a good pair of coefficients that does lead to the learning of the training set is found, a better set can exist that will lead to much faster learning. It is frequently desirable to change the values of the coefficients after the net has gone through an initial period of learning. This usually takes the form of

experiments. It failed, however, to distinguish the three pairs of identical individuals mentioned here.

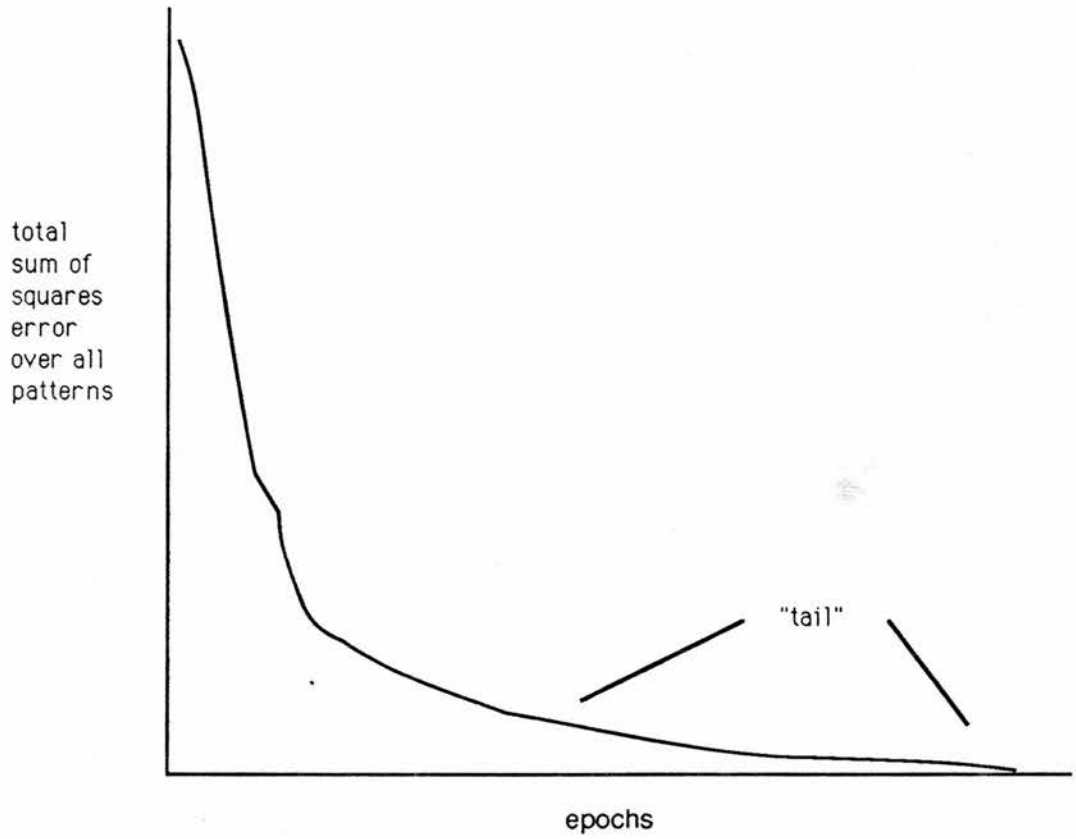


Figure 5.5: A typical learning curve

increasing the value of η and α once the error curve has reached its long 'tail' (see Figure 5.5).

A promising idea that was aimed at automatically adjusting the values of η and α during learning is the adaptive training algorithm reported by Chan and Fallside (1987). They used the angle between the current error gradient vector and the previous weight update vector to give some idea of the local shape of the energy surface. If a ravine is detected, η is reduced to avoid oscillation across the sides of the ravine. If a plateau is detected, η is increased to counteract the effect of the small gradient. α is made to be proportional to η depending on the ratio of current error gradient vector magnitude and past weight update vector magnitude. Their algorithm appears to work well for a vowel recognition problem and an image recognition problem that they describe, choosing parameters that are close to

the optimal ones. However, the method, at least the implementation of it that I attempted, lead to very unstable learning for the networks used here and conferred no advantage, often leading to curves that didn't converge.

For the simulations reported in this chapter, the values of η and α were nearly always fixed at 0.05 and 0.9 respectively. This method was convenient for the many repeated simulation runs that were needed. These parameters appeared to be the most suitable for the initial phase of learning and it was judged more convenient to let the simulations run to completion rather than to continually examine them and change the parameters in an attempt to speed learning.

The initial values of the weights were fixed at small random values between the limits of -0.25 and $+0.25$.

The problem of 'stuck' output units

A more serious problem with the back-propagation algorithm as far as the training data used in this research is concerned is the way in which the activation of output units can become 'stuck' at values approaching 0.0 or 1.0 for some of the patterns in the training set. This occurs because of the way the error at the output is multiplied by the derivative of the logistic activation function. For an output unit, j,

$$error_j = t_j - o_j$$

where t_j is the target activation and o_j is the actual activation of j. The total error signal, δ_j is given by:

$$\delta_j = error_j o_j (1 - o_j)$$

So, as the unit's output approaches 1 or 0, the amount of error back-propagated

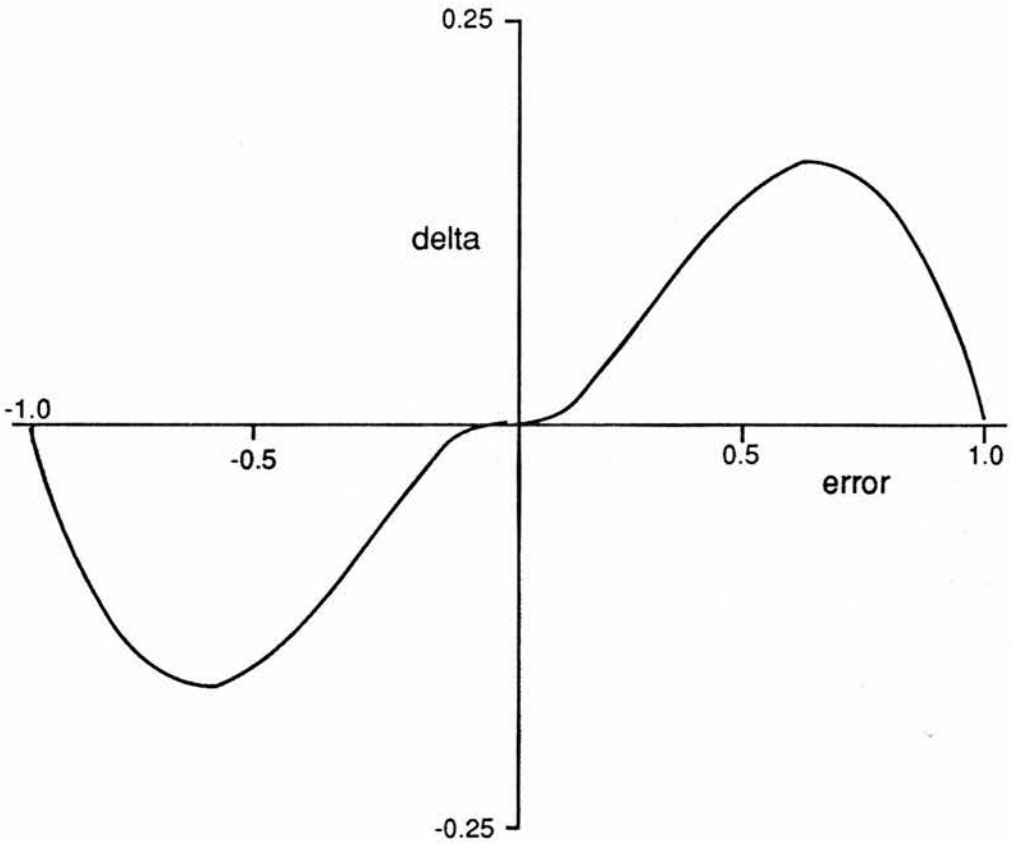


Figure 5.6: The standard error curve

approaches zero, and so the degree of weight change tends to zero since the change in a weight between a unit, i and a unit j, Δw_{ij}, is given by:

$$\Delta w_{ij} = \eta \delta_j o_i$$

where η is the learning rate coefficient. The situation is illustrated in Figure 5.6 where δ is plotted against the error. The error signal is zero when the error is zero but it is also zero where there is a maximal error of $+1$ or -1 because o_j will be 0 or 1 in this situation which leads to a $o_j(1 - o_j)$ term of 0. Thus, if there is a situation in a learning trial where the error (target - output) approaches 1 or -1, the unit may 'lock up'. Theoretically, the unit should recover eventually but this may never happen because of rounding error [2]. This phenomenon does not always lead to irrevocably 'stuck' units but can decrease the speed at which a

[2]There will be a point where the output of the unit is represented by $0.0\cdots0$ or $1.0\cdots0$

particular pattern set is learnt. For some starting points, the networks used here were badly affected by this problem, especially the ones in Section 5.5. Rather than repeat the simulations with different random starting weights until a good learning trajectory was found, I attempted to alter the learning algorithm slightly to circumvent the problem.

Fahlman (1988) points out this problem and suggests a very simple alternative. He suggests that 0.1 is added to the above expression so that its lowest value is 0.1 [3]. Fahlman shows that this solution speeds the learning in the small encoder-decoder networks that he uses as benchmarks. Unfortunately, this rather unprincipled method fails on more complex networks and leads to very unstable learning in the networks used here.

In an attempt to keep the beneficial effects of Fahlman's method (preventing the error signal from reaching zero) but to avoid disrupting the course of normal learning, I modified his scheme so that the increment only came into effect when the output of the unit was below 0.1 or above 0.9. This performed somewhat better but was still not satisfactory.

A better solution was suggested by Lionel Tarassenko (personal communication), and was adopted for the work done here. He reasoned that the shape of the error signal function should be a straight line, so that the signal is at its greatest when the error is at its greatest. His scheme with a slope of 0.25 is, in terms of pseudo-code, where $error = \delta_j$:

```
If (error < 0.0)
    then error = -0.25/(1.0 + error)

else error = 0.25/(1.0 - error)
```

because of the limited resolution of the representation of a floating point number on a machine with a fixed number of bits used to represent each number. When this happens, the error signal reaches zero and no further learning will occur.

[3] In this and all the other alterations to the standard back-propagation learning algorithm described in this section, the changes are made to the error expressions for the output units only.

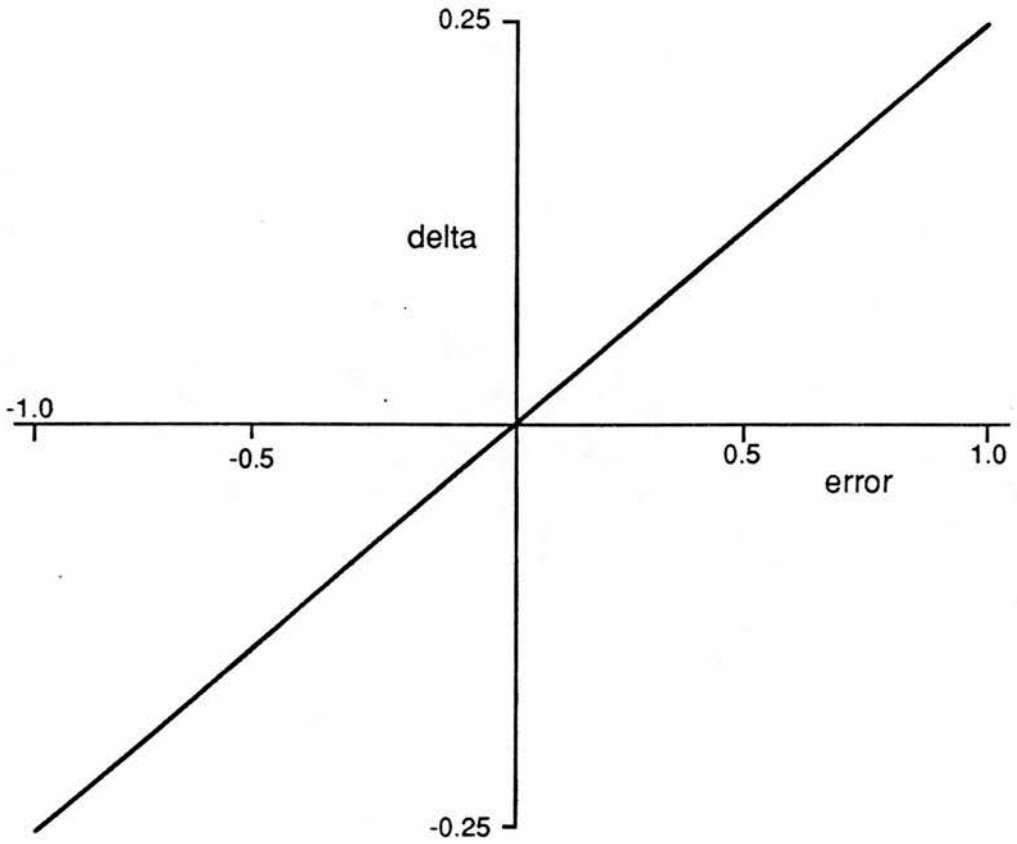


Figure 5.7: Tarrasenko's linear error curve

The function is plotted in Figure 5.7. He has shown that this method speeds the learning of simple encoder-decoder networks more than Fahlman's method. What was more important for the research reported here was that it would lead to stable learning in my more complex networks. On the whole, the method was satisfactory in preventing output unit lock-up. What was substituted was a very long 'tail' in the learning curve, leading to very long learning times while a few patterns had their rogue units 'pushed' down from errors close to 1 or -1 while the other patterns had already reached criterion. The method did not produce perfectly smooth learning curves but caused short-lived spikes of increased error.

5.3.3 Simulation

Once the network has been trained, it has been demonstrated that such a system can be constructed to perform the inference necessary to produce a well formed property specification from a consistent input feature vector. The next stage is to disrupt the network to see whether errors can be generated and whether the distribution of errors is similar to that of human subjects. The disruption is caused by injecting noise into the input layer. A particular chance of 'flipping' any particular input node is chosen and each of the input nodes has that probability of having its value changed from 0 to 1, or from 1 to 0 for each run of the network. During the simulation process, the weights are fixed and no learning is done. The set of patterns run through the network during the simulation is generated from the 1537 paragraphs actually given to the original subjects.

When the input vector is disrupted, it is likely that the net will be producing an output vector from an input vector that it was not trained on. This could make the activations of the output units stray from values close to 0 and 1. A thresholding mechanism was imposed so that any value of activation of an output unit less than 0.5 was regarded as 0 and an activation of 0.5 or higher is regarded as 1. This is necessary to compare the responses of the network to the psychological data. It is at this simulation stage that we can discover whether the network is capable of error correction, inconsistency resolution and production of an error distribution similar to that in the human data.

The somewhat crude form of disruption used here was deliberately chosen as the simplest method of inducing errors.

Each trained network was used as the basis of two simulations of the production of recall errors. The simulations differed by the seed that was used for the random numbers that produced the disruptive noise that was injected into the input layer. The recall category frequencies were transformed in the same way as those obtained from the human subjects, by dividing by chance opportunity and logging. A correlation coefficient between the logged and adjusted human data and

the logged and adjusted simulated data was then calculated as a rough measure of the degree of congruity of the simulations with the data obtained from human subjects. The coefficient is squared to obtain the R^2 statistic or proportion of total variance accounted for by the correlation.

It was deemed more important that the networks should produce the theoretically interesting multiple errors than have an exact fit with the human data. As explained earlier an exact fit with the data was unlikely due to the simplifying assumptions used; the philosophy being to reproduce the main effects with as few assumptions as possible.

5.3.4 Generalisation

One of the attractive properties of PDP systems is their ability to generalise. Good generalisation shows that a network has learnt what we would consider to be the characteristics of the training set rather than storing the information in an uninteresting 'table look-up' manner. The usual way of demonstrating any generalisation ability in a PDP network is to test the trained network with some input patterns from the same domain as the training set but ones that the network has not been trained on. For the networks described here, this technique is impossible to employ because the training set contains all possible input vectors. The method used here was to remove a fraction of the training set, train a network, and use the missing vectors as a test of generalisation behaviour. These test networks were only used as a way of investigating the likely representation of the complete network and were not used to simulate recall error.

5.4 Performance of the First Network

The first network was designed to be the simplest possible PDP implementation that could compute a mapping from a distributed feature representation based on the statistical recall error model to a simple specification of the 8 vocabulary

items of a recall. Once trained on all the possible input vectors, each network was used to simulate the production of recall errors. Each network was also subjected to two tests for its ability to generalise.

This section will describe the architecture of the networks and the way they were trained. The results of the simulation and generalisation test phases will also be discussed.

5.4.1 Architecture of the network

The network consisted of three layers (see Figure 5.8). The input layer consisted of 15 units corresponding to the variables extracted by the multiple regression. The middle layer consisted of 14 units. This number was chosen as a result of past experience as a compromise between a smaller number of units that would cause long learning times and a larger number of units that might adversely affect the generalisation abilities of the network. The output layer consisted of eight units.

5.4.2 Training

The training set consisted of 133 patterns – one of each possible unordered pair of individuals minus the three vectors that could not be distinguished by the feature set.

Ten trials were run, each with a different random starting point in weight space identified by an arbitrarily chosen seed for the random number generator. The trial was stopped when there were no errors on any of the output units for any of the training patterns greater than 0.2. The number of epochs needed to reach this criterion varied from 210 to 490, with a mean of 336. Table 5.1 shows the relevant statistics for the 10 networks for training, simulation and generalisation tests. Each network is identified by the the seed which was used to generate its initial random weights. The table lists the number of epochs needed to train the net, the R^2 statistics for both simulations, and the epochs needed for training and

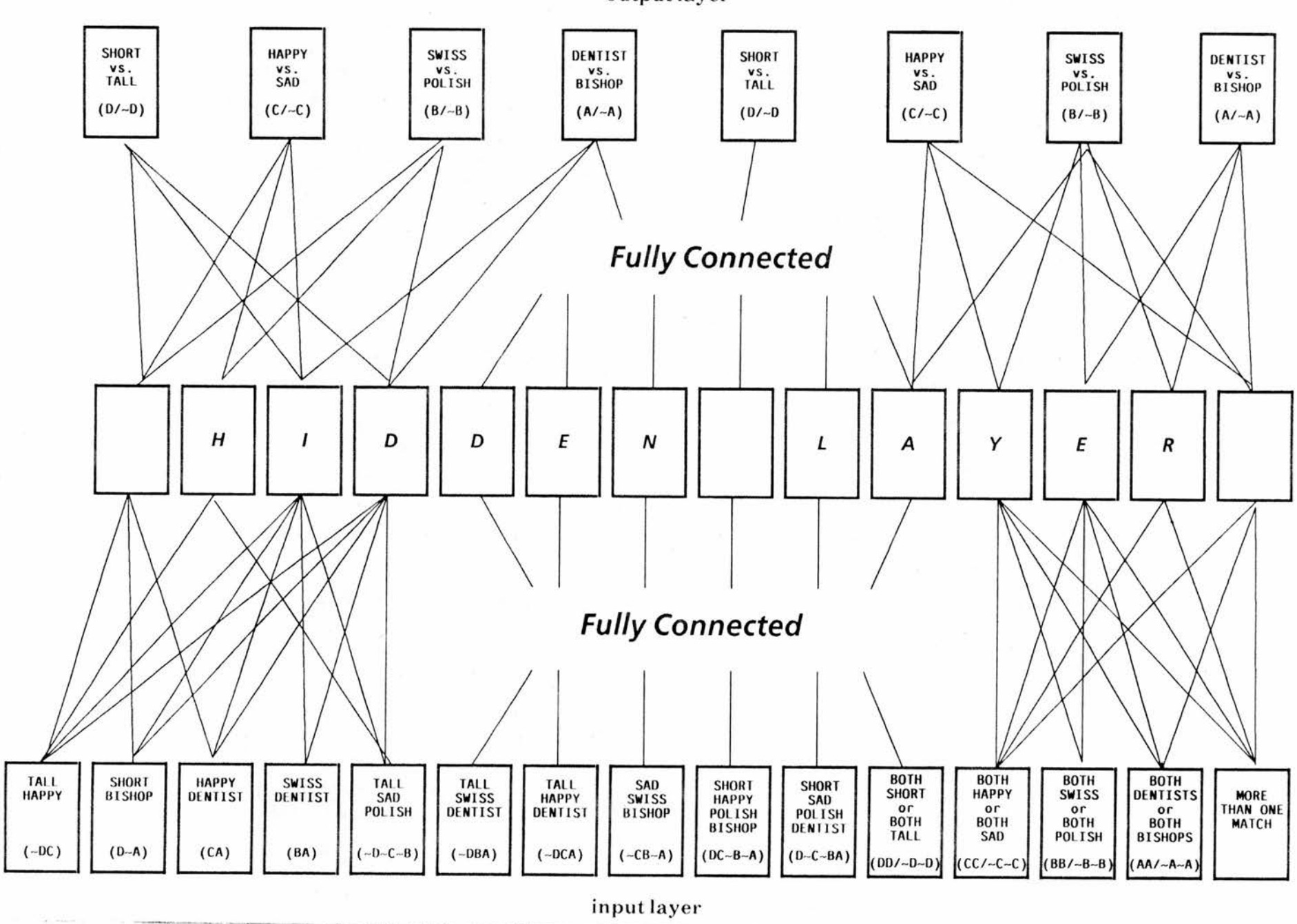


Figure 5.8: The First Network

	Networks labelled by their randomising seed										
	227	383	386	525	598	627	658	678	704	989	Mean
epochs to learn	377	451	242	310	490	332	210	459	220	272	336
R^2 no seed	0.69	0.69	0.68	0.62	0.70	0.69	0.72	0.68	0.69	0.70	0.69
R^2 seed 1210	0.68	0.71	0.70	0.67	0.71	0.69	0.69	0.70	0.73	0.67	0.70
Test a:											
epochs	255	398	217	265	252	227	189	300	419	249	277
vectors (/12)	6	7	6	6	4	4	7	7	7	7	6.1
units (/96)	85	88	87	86	84	85	88	89	88	89	87
Test b:											
epochs	338	303	422	356	244	306	243	516	273	543	354
vectors (/12)	4	5	9	7	8	8	7	7	9	6	7
units (/96)	86	86	91	91	90	92	91	90	90	89	90

Table 5.1: Summary table for the training, simulation and generalisation statistics of the first network type. The number of epochs each of the ten networks required to learn the training set is given as well as the R^2 values for the errors produced on both simulation runs. The numbers of epochs and accuracy in terms of vectors and units are given for both of the generalisation tests.

the measures of success for both of the generalisation tests.

5.4.3 Simulation of Recall Error

Each of the 10 training trials was used as a basis of two simulations using 3% noise. The simulations are referred to by the random seed used in their training followed by either an A or a B for the two simulations. All the A simulations used the same noisy input as did all the B simulations. The R^2 statistic was similar for all 20 simulations (see Table 5.1), ranging from 0.62 to 0.73. Thus the networks account for a large proportion of variance in the data.

To get a better idea of how similar the error patterns of the network simulations were to the human data, seven statistics derived from the error frequencies were compared (see Table 5.2).

The first two numbers compared were the total frequency of correct responses and miscellaneous errors. These mark out the two extremes of recall accuracy and opportunity for making a response. On the one hand a correct response is totally correct while a miscellaneous error is likely to be very inaccurate; on the

Net	corr	misc	sgl/2	ipol	isl/2	dc/dh	ppol
human	1087	30	0.62	101	0.45	2.47	18
227A	1106	51	0.72	37	1.81	0.74	13
227B	1047	44	0.97	33	1.12	0.81	10
338A	1166	30	0.92	50	1.36	0.86	15
338B	1127	30	0.87	66	0.95	1.52	15
386A	1146	29	0.82	49	2.2	1.3	12
386B	1098	27	0.98	47	4.7	1.47	14
525A	1191	26	1.08	45	1.9	1.4	12
525B	1145	24	1.1	52	1.54	0.82	9
598A	1140	40	0.86	32	0.83	0.77	10
598B	1084	44	1.20	35	1.79	0.91	12
627A	1132	28	1.43	42	2.1	0.92	16
627B	1126	38	1.47	50	2.18	0.85	13
658A	1133	36	0.65	51	1.67	0.89	10
658B	1116	31	0.74	55	1.63	1.0	8
678A	1136	33	1.0	26	0.93	1.3	14
678B	1095	39	0.84	40	1.64	1.43	12
704A	1131	30	0.93	31	1.56	0.87	17
704B	1097	28	0.92	41	1.64	1.10	12
989A	1105	56	0.85	28	1.56	0.86	12
989B	1058	63	1.16	29	2.72	0.71	12
Mean	1119	36	0.98	42	1.79	1.03	12
S.D.	34	11	0.22	11	0.83	0.27	2

Table 5.2: Summary table of some of the errors for the first network type. Seven error frequency statistics are given for each of the 20 simulations as well as for the human data.

other hand there is only one correct recall for any description but there are very many ways of making a miscellaneous response. The third statistic is the ratio of the frequency of single errors on the first individual to single errors on the second individual. This order dependent effect was quite striking in the data and was one of the factors leading to the second network described in this chapter. The fourth number was the total frequency of individual polarity errors, another striking aspect of the data. The fifth statistic was another order dependent one, the ratio of individual polarity + singleton errors on the first individual to those on the second individual. The sixth statistic was the ratio of double complementary errors to double homogeneous errors, one of the reasons for proposing the nmat feature, since although both are double errors, double complementary errors preserve the truth value of nmat while double homogeneous change it. The final statistic compared was the total frequency of property polarity errors, another important class of double error.

Table 5.2 compares the values of the seven numbers for the human data with the values for the 20 different simulation trials. The means and standard deviations of the simulated data are also shown. It is not surprising that the frequency of correct responses matches fairly well since it was on that basis that the degree of random noise used to induce errors was chosen. There are slightly too many correct responses from the nets, with a low standard deviation. The match for frequency of miscellaneous responses was reasonably good. This is quite important since it shows the nets making a small but appreciable number of severe errors, even though most responses are correct. This is good evidence for underlying structure in the representation used by the network. The ratio of single errors on the first individual to those on the second is not simulated by the networks. The other order dependent statistic, the ratio of individual polarity + singleton is not simulated either. The frequencies of individual polarities and property polarities are lower for the networks than in the data but still respectable. The networks fail to simulate the greater number of double complementary than double homogeneous errors in the human data.

χ^2 tests were performed comparing the distribution of human recall errors with

those of the simulations. The data was collapsed over matchtype to avoid problems of sparsity and empty cells. In every case, the χ^2 statistic was highly significant, indicating that the distributions are different. Thus, although there is a good correlation between human and simulation recall frequencies, the simulations are not good enough to produce a similar overall recall error distribution to that of the human data. Z-tests were then performed, comparing the mean frequency of each human recall error category with the mean and standard deviation of the corresponding categories in the simulations. In 14 out of 20 cases at the 5%(2-tailed) level, the tests suggested that the human data points were part of the same distribution as the simulation data points. For 6 of the error categories (sg1-, sg2+, ipol, 2cs2, dhs2, mirr) this was not the case. The Z-tests are more encouraging than the χ^2 tests but it is clear that the simulations are not yet fully satisfactory.

Although not a perfect match, all the networks turn in a respectable performance. They compare very well to ten simulations where the eight recall properties themselves are subjected to 3% noise and processed by the recall scoring algorithm (see Table 5.3). As would be expected for a representation with no underlying structure and no opportunity for error correction, the noise causes errors with no evidence of dependencies between properties – multiple errors are much more unlikely to occur than single errors, and the more severe a category of error is, the less frequent it will be.

These 'null criterion' simulations produce more correct responses, but the networks do not compare too badly when it is remembered that they are subjected to 3% noise on 15 units rather than just 8, and hence have a much greater degree of disruption. Table 5.3 shows very few of the severe miscellaneous errors and very low frequencies of the important individual and property polarity errors. The ratio of first to second individual single errors is about the same as that produced by the network. The mean ratio of double complementary to double homogeneous is a little better than the result from the network simulation but the standard deviation is high. There are too few of the triple individual polarity + singleton errors to calculate a first to second individual ratio.

Net	corr	misc	sgl/2	ipol	isl/2	dc/dh	ppol
human	1087	30	0.62	101	0.45	2.47	18
1	1200	1	0.89	3	-	1.89	1
2	1224	1	1.08	1	-	1.29	1
3	1211	2	0.99	6	-	1.88	1
4	1199	0	0.76	1	-	2.11	1
5	1200	0	0.94	4	1/0	0.94	0
6	1198	2	0.93	1	-	1.00	0
7	1224	0	0.98	4	-	1.89	2
8	1196	1	1.08	4	-	1.29	2
9	1224	0	1.01	2	1/0	0.58	1
10	1189	1	0.98	0	1/0	1.78	1
Mean	1207	0.8	0.96	2.6	-	1.47	1
S.D.	13.2	0.79	0.09	1.9	-	0.51	0.66

Table 5.3: Summary table of errors in random noise simulations

In conclusion, we can say that the network simulation has accounted for a large proportion of the variance in the human data. It produces a good but not perfect match of many of the frequencies of many of the recall error categories. Its performance compares very well with a simulation of noise applied directly to the output properties. The network simulations were unsuccessful in matching some of the more subtle aspects of the data including serial order effects and the ratio of double complementary to double homogeneous errors.

5.4.4 Generalisation behaviour

Two different generalisation training sets (set a and set b) were made up by removing 12 vectors from the complete training set. These were arbitrarily chosen with the constraints that none described identical individuals and only half had matched introducing dimensions. Descriptions of identical individuals were felt to be particularly unrepresentative of the stimuli that subjects were given and it was felt desirable that at least half of the test set should contain representations of the non-matched introducer descriptions that subjects received as stimuli. The 12 vectors removed represented 10% of the members of the training set that didn't describe identical individuals. These training sets are designed to test whether the architecture chosen for learning the complete training set is likely to be ex-

tracting some generalisations in the mapping of feature truth values to property specifications. All the generalisation tests were started from the same position in weight space (i.e. using the same seed for the random number generator that controlled the values of the initial weights) as their respective complete learning trials.

A correct output unit is taken to be one that has an error of less than 0.5. Two measures of successful generalisation are used. The first and most stringent is the proportion of output vectors from the test set that are completely correct. The second is the proportion of correct output units in the entire test set. It can be seen from Table 5.1 that both tests behave respectably, the number of complete output vectors being correct ranging from 4 to 9 out of 12 and the number of complete output units correct ranging from 84 to 92 out of 96. These successful generalisation tests, though not perfect, are good evidence that the representation encoded by the weights and biases in the networks trained on the complete training set has not 'memorised' each input vector to output vector mapping but has extracted some generalisations between similar input-output pairs.

5.5 Performance of the Second Network

This network was designed to extend the model by enabling the network to map one feature value vector to both possible orders of recalled individuals. This was done by adding an extra unit to the input layer which made the training of the second type of network harder than for the first network. This section will describe training, simulation and generalisation tests for the second network in a directly comparable way to the treatment of the first network.

5.5.1 Architecture of the Network

The previous network is successful at learning how to make a correct response when given a well-formed input vector and is able to produce the same sorts of multiple errors as those found in the data. However, the 'order of recall' of the individuals in the output layer was fixed, i.e., a vector of feature truth values produced only one of the two possible recall orders of individuals that it specified. In the experiment itself, subjects were not constrained to make their recall in any particular order (see discussion in Chapter 3), and so this is one aspect of the human performance that the first network is not capable of simulating. What is needed is some mechanism whereby the same feature vector can specify two possible output vectors which differ in the order of their specification of the two individuals (see Figure 5.9)

As it stands this requirement is impossible to achieve within the chosen back-propagation framework. Since all weight changes during training depend on a difference between the target and actual output vectors produced from one of the input vectors in the training set, the same input vector cannot be trained to be associated with more than one output vector. To get round this problem an extra unit was added to the input vector - the so-called 'cueing unit'. The activation of this unit served to control the order in which the two individuals specified by the other input units appeared on the output vector (see Figure 5.9). This extra unit served as a minimal contextual cue to distinguish a recall of, for example, 'a short fat Polish bishop and a tall fat Swiss dentist' from 'a tall fat Swiss dentist and a short fat Polish bishop'.

As well as allowing the two different recall orders to be produced from the same feature truth value vector, it was hoped that this additional constraint would increase the degree with which the net tended to represent the input vector information as describing two individuals rather than just eight properties. One way of testing whether this strategy is successful is described in Section 5.5.4. It consists of tests of how well the net can generalise to this task.

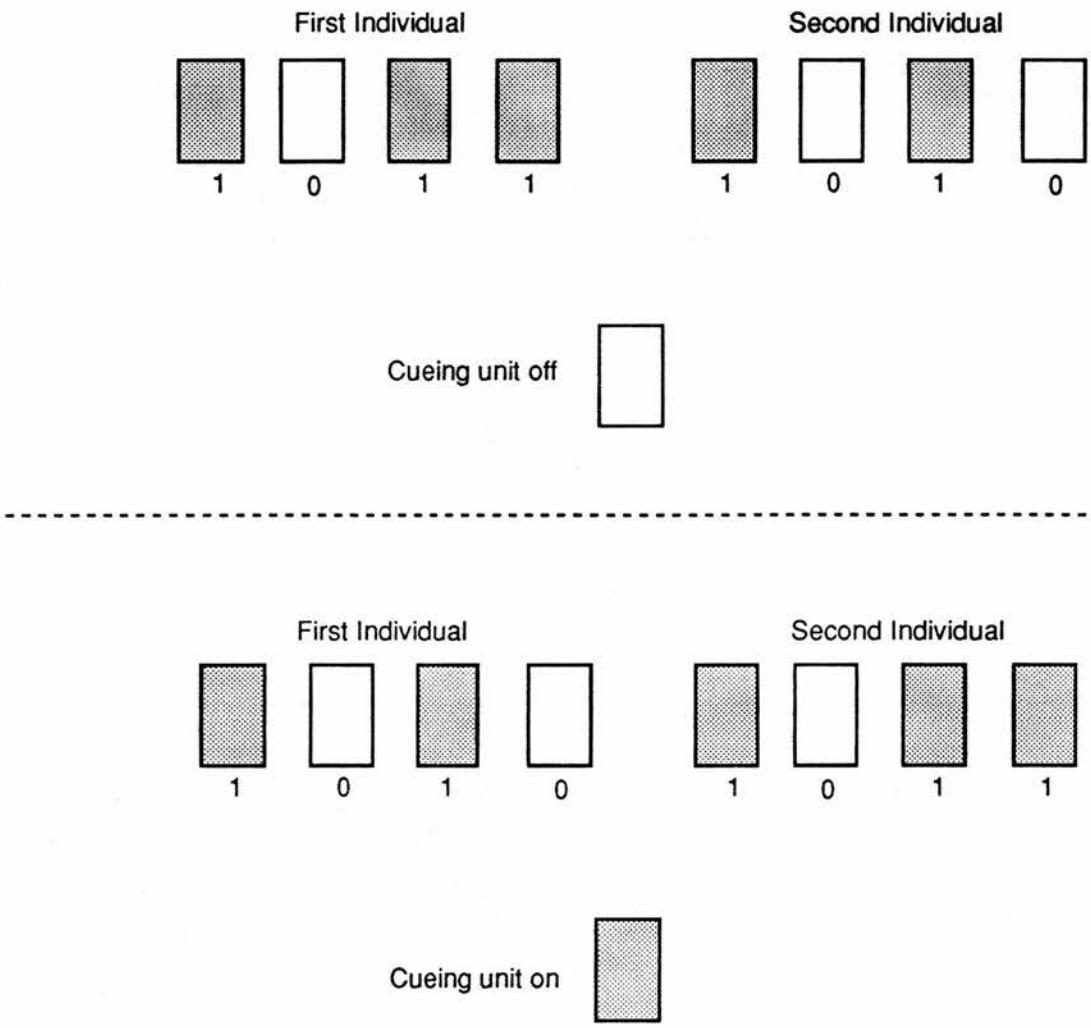


Figure 5.9: An example of the use of the cueing unit. The position of the two recalled individuals is swapped when the value of the cueing unit changes.

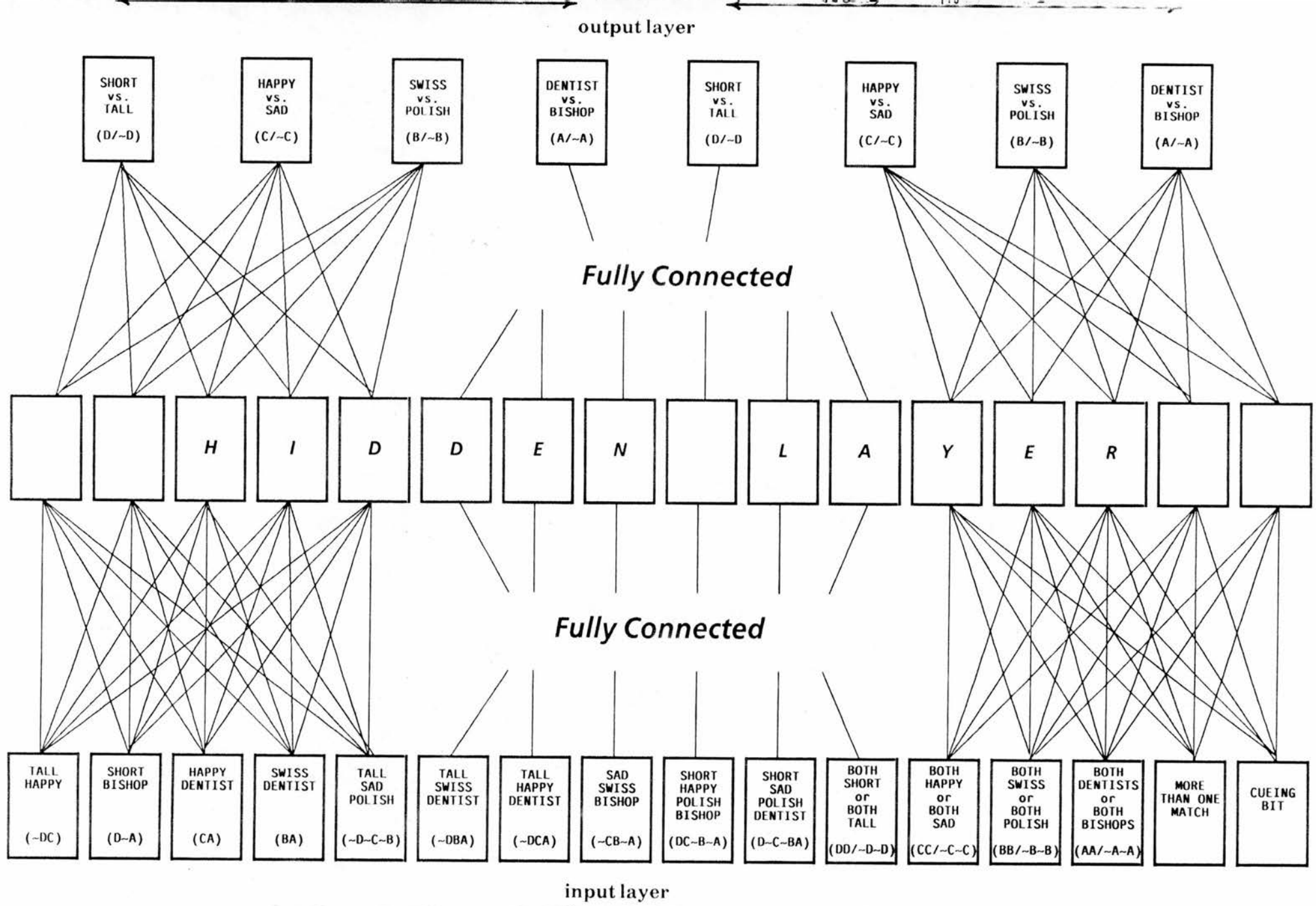


Figure 5.10: The Second Network

The input layer contains an extra unit, making 16 altogether. The hidden layer was enlarged to 16 units because the additional constraint of making the network output two output vectors from input vectors that only differ by the activation of the last input node, makes the task of the network more difficult. As before, the output layer has 8 units.

5.5.2 Training

The training set now contains 253 members — all possible combinations of values of the output units minus the three that cannot be distinguished by the feature set. The number of extra members compared with the first network are merely the extra 120 created by allowing both individual recall orders to be represented by means of the cueing unit.

The cueing unit was set so that it had an activation of 0 if the 4 bit number made up from the activation of the output units representing the first individual was greater than the 4 bit number made up from the units representing the second individual. If the first 4 bit number was less than or equal (i.e. a description of identical individuals) to the second 4 bit number, then the cueing unit was set to have an activation of 1. This rather arbitrary way of setting the cueing units was done so that the only 'meaning' of changing the activation of the cueing unit was to change the order of the output individuals. As well as learning the logic relating the input features to the output properties the network also had to learn that the cueing unit controlled the order in which the two individuals were 'recalled'.

The extra task required of the network proved to be very taxing to learn. Out of the 16 different random initial configurations tried, only 7 converged in a reasonable amount of time. The results reported are based on these 7.

The problem of the tendency of output units to have errors of 1.0 or -1.0 for a few units on a few of the members of the training set is particularly apparent for these training sets. The modified training algorithm was able to cope with this but only at the expense of extremely long 'tails' in the learning curves of some

	028	049	069	084	096	633	650	Mean
epochs to learn	1851	1448	5207	300	844	6394	917	1959
R^2 no seed	0.69	0.70	0.68	0.65	0.65	0.76	0.74	0.70
R^2 seed 1210	0.74	0.74	0.74	0.63	0.69	0.73	0.71	0.71
Test a								
epochs	426	972	4960	879	6000*	773	553	2080
vectors (/24)	22	19	10	21	12	15	12	15.9
units (/192)	188	185	170	187	174	170	170	177.7
Test b								
epochs	405	638	2270	2843	300	1194	341	1142
vectors (/24)	14	9	14	14	11	11	16	12.7
units (/192)	175	166	178	175	171	170	177	173.1
Test c								
epochs	2935	895	6000*	354	3551	606	701	2149
vectors (/12)	6	10	7	8	8	9	8	8
units (/96)	87	93	87	88	85	91	92	89
Test d								
epochs	3000*	983	6000*	322	720	7484	235	2678
vectors (/12)	4	4	5	5	9	7	4	5.4
units (/96)	85	84	86	86	92	84	83	85.7

*didn't learn to criterion but tested anyway

Table 5.4: Summary table for the training, simulation and generalisation statistics of the second network type

of the trials. The number of epochs needed to learn the training set to criterion ranged widely from 300 to 6394, with a mean of 1959 (see Table 5.4). The longer training periods were due to the extremely long tails in the learning curves caused by a minority of the members in the training set. It might perhaps have been more convenient to have used a greater number of hidden units since this might have decreased the number of epochs needed to reach criterion. However, since a reasonable number of networks did converge within a practical time, and since a constrained hidden layer can only add to the generalisation ability of the network, the choice of 16 hidden units is probably vindicated.

5.5.3 Simulation of Recall Error

Again, each training trial was used as the basis for two simulations of the production of recall errors using 3% noise. The R^2 values ranged from 0.63 to 0.74, indicating that a very good proportion of variance has been accounted for.

Net	corr	misc	sgl/2	ipol	isl/2	dc/df	ppol
human	1087	30	0.62	101	0.45	2.47	18
028A	1139	28	1.1	68	2.27	0.9	36
028B	1131	22	1.1	82	0.67	1.2	29
049A	1163	25	1.27	27	1.1	1.4	36
049B	1130	19	1.07	41	0.64	0.5	34
069A	1128	21	1.01	48	1.8	0.48	30
069B	1126	31	0.98	52	0.91	0.9	24
084A	1219	8	1.02	49	0.57	0.8	28
084B	1176	6	1.2	66	0.24	0.56	28
096A	1191	10	1.1	55	0.78	1.27	33
096B	1140	8	1.25	66	0.76	1.05	25
633A	1125	23	1.15	43	5.3	1.16	23
633B	1090	23	1.16	61	2.5	0.68	15
650A	1139	30	0.91	51	0.94	1.0	29
650B	1100	27	0.81	70	1.04	0.88	37
Mean	1143	20	1.08	56	1.39	0.91	29
S.D.	34	9	0.13	14	1.30	0.29	6

Table 5.5: Summary Table of Errors in the Second Network Type

As for the previous network type, seven summary statistics were examined for each network. Table 5.5 should be compared with Table 5.2. As far as these summary statistics are concerned, there is little difference between the two network types. The new networks produced more correct responses, individual polarity errors and property polarity errors but less miscellaneous errors. There is little difference between the more subtle ratio statistics. It is particularly disappointing that the new networks are no better at the serial order related ratios than the older networks.

As in the first PDP model, χ^2 tests showed that the overall error distributions of each simulation were significantly different to the human error distribution. Z-tests showed 5 error categories (sgl-, ipol, 2cs2, dhs2, mirr) where the human data points did not seem to be part of the same distribution as the simulation points. Again, it is clear that the simulations do not perform satisfactorily on some of the error categories.

5.5.4 Generalisation behaviour

Two different types of generalisation test were performed. The first (Tests a and b) was analogous to the ones used for the previous network. However, instead of 12 single members of the training set being removed, 12 *pairs* were removed, the members of the pairs only differing by the activation of the cueing node. The pairs removed consisted of the same feature values as the single members removed in the previous generalisation tests. These pairs again represented 10% of the training set (excluding identical individuals). The tests (see Table 5.4) demonstrated a similar degree of generalisation for the mapping of feature vectors onto property specifications as shown in the first network. The degree of success of the 14 tests varied widely, ranging from 9 to 22 completely correct vectors out of 24 and 166 to 188 correct output units out of 192. One of the learning curves for the test training sets (seed 096) for Test a did not fully converge, but its weights were used for a generalisation test anyway. Its performance was satisfactory.

The second type of generalisation test (Tests c and d) was designed to test generalisation for the ability to output both output individual orders from the same feature value vector, i.e the ability to swap the order of output individuals when the activation of the cueing unit is changed from 1 to 0 or from 0 to 1. These tests removed 12 single members from the complete training set, each members of different pairs whose input vectors differed only by the value of the activation of the cueing unit. The 12 members represented a test of 10% of the pairs in the complete training set (excluding identical individuals). Six of the members represented descriptions with matched introducers and six represented descriptions with mismatched introducers. Within each of these groups of six, three had an activation of 0 on the cueing unit and three had an activation of 1.

Test c was consistently easier than test d, with a mean of 8 complete vectors correct out of 12 and 89 correct output units out of 92 as opposed to 5.4 correct vectors and 85.7 correct output units. All of the tests achieved a respectable degree of generalisation including the three whose learning curves did not converge.

5.6 Summary of PDP Model

The PDP model of recall is an extension of the statistical model of recall errors and as such embodies many of the same hypotheses and shares the same contrasts with previous work. The chief methodological success of this model is that, by specifying the architecture of a network directly from a statistical analysis of real data, we depart from much of the work on connectionist modelling of psychological phenomena where an arbitrary network is chosen in the less well-founded hope it will simulate the data. An important aspect of the model is the way it focusses our attention on the issues of inference from a distributed database and the resolution of inconsistencies when the database is disrupted. In this model the inference is from a set of feature truth values to a correct recall. Inconsistency resolution takes place when the network produces a recall, whether correct or not, when one or more feature values have been changed in such a way that the set of feature values as a whole is inconsistent.

5.7 The Usefulness of the Approach

This section will discuss how successful the use of this modelling framework has been at extending the recall error model. It will also compare the two types of networks that were simulated.

The networks described in this chapter have been successful at usefully extending the statistical model described in Chapter 3. They show that a highly simplified system such as a three-layer net with flat random noise injected into its inputs is capable of inconsistency resolution. Furthermore, the consistent outputs produced were a good simulation of the types of responses produced by human subjects. The incorrect responses fell into the error categories seen in the human data, reflecting the underlying dependent structure that had, after all, been extracted from the human data by the statistical analysis. The fact that such a simple system can perform the task of inconsistency resolution and produce a similar

error distribution to the human data supports the theoretical interpretation of the statistical model, since it demonstrates the existence of a feasible computational mechanism for such processes.

The second network was a refinement of the process model in that it was capable of producing both output individual orders from essentially the same input feature truth value vector. This ability did not make it simulate the human data any better but it did show how flexible the basic network architecture was. The extra functional ability conferred by the cueing unit mechanism is a significant enhancement to the model because it simulates a more realistic, less static way of recalling a pair of individuals. Any extra dynamic ability would require a different architecture since layered feed-forward networks trained using back-propagation do not allow any feedback or relaxation. The reason that such an architecture was used in this research was that it was the most practical way to learn the complex inferences required.

The back-propagation network architecture used here was a rewarding and flexible modelling environment to use. The networks proved capable of learning the required input-output associations and demonstrated excellent generalisation capabilities. Although the simulations did not match the data exactly, they were impressively close for systems with such a low number of parameters. Possibilities for extending the research done using these networks is discussed in Chapter 6.

Chapter 6

Conclusions and Further Work

This final chapter will summarise the work done, evaluate its successes and weaknesses, and suggest the areas where further work might prove fruitful. A brief description of finished and ongoing work that makes use of the work described here will be given.

6.1 Summary

The aim of the work described in this thesis was to develop research methods for the study and modelling of the way in which humans represent the attributions of properties to individuals. Despite its neglect in the literature, this seemingly trivial problem has been shown to demand a considerable amount of cognitive resources. As the previous chapters have demonstrated, it has been amenable to experimental investigation and modelling. The experimental and modelling methods that have been developed have allowed considerable progress in our understanding of this phenomenon and will provide a basis for further work.

The experimental paradigm chosen has allowed the collection of data susceptible to analysis and modelling. Unusually, we were able to collect both reading time data reflecting the constructive processes building up a representation and recall error data that could be used to infer structural aspects of the representation. The

observation that sparked this whole project, the semantic ordinal effect (Stenning 1986) proved general to a wide range of experimental manipulations: e.g. subjects, experiments, semantic domain, text organisation, recall task, disruptor task. The MIT has proved to be extremely versatile, allowing the large variety of different manipulations to be carried out within the same paradigm.

The control of matchtype in the Antonymy Experiment revealed that the matching and mismatching relations between individuals has a major effect on reading time. This observation led to the proposal of the different load variables, MISLOAD, MATLOAD, NEUTLOAD and LOCMIS that accounted for much of the variance in the reading time data from the Antonymy and Replication experiments. The theoretical interpretation of these models was that the semantic ordinal effect could be explained as the increasing effort needed to recruit the associations necessary to support the attribution of the correct properties to the correct individual. This effort is more difficult for mismatched dimensions since they make the attribute binding problem harder by increasing the number of possible distinctions. The regression models were both satisfactory for the proportion of variance they accounted for and for the theoretical interpretation that could be placed on them.

The large amount of recall error data generated by the Replication Experiment allowed a taxonomy of different error classes to be made. It was clear that the matching and mismatching of vocabulary dimensions was important in the structure of the underlying representation since many errors appeared to involve the correct memory for the matching or mismatching of a dimension but the forgetting of the assignment of properties to individuals. The relatively high frequencies of some error categories, the most common example being the individual polarity error, appeared to be manifestations of redundancy in the dependant structure of the representation — *both* intra-individual *and* inter-individual associations were being made.

The important assumption that a stimulus would be more likely to be confused with another description the more similar that description was to the stimulus,

coupled with the proposal that the representation could be modelled by independent redundant features allowed the different frequencies of recall error categories to be modelled using multiple regression. For every error category, a vector of the mean number of disruptions caused for a number of candidate features was calculated and used to find out which features were the most successful in accounting for the log adjusted frequencies of the error categories. The less feature disruption an error caused the more likely it was to occur. Effectively, the multiple regression procedure extracted a similarity metric between different descriptions in terms of the numbers of feature values they share and the salience or importance of the features. The more similar a potential recall was to a stimulus the less disruption was needed to cause that response and the more likely the response was to occur.

The first regression model of recall errors used intra-individual features that were tagged according to the individual involved. The statistical model successfully described the data but did not explain *how* properties are assigned to individuals. The preferred model was based on features that were simply true or false of the description and had no referential tags. This model accounted for the same amount of variance as the first model but assumed the existence of an extra process that was capable of inferring which features were true of which individual. Thus, although the second model did not explain how this process would work, it provided the impetus for the next stage of modelling. The second recall model proved to be a particularly convenient basis for a PDP network.

The multiple regression model itself allowed the specification of a PDP network model of recall processes. Networks were trained to output the eight properties of a description when given an input vector corresponding to the truth values of the features from the regression model. The network simulation was refined so that it was capable of producing both orders of output individuals from the same pattern of feature truth values. The redundancy of the input layer representation specified by the regression model meant that any disruption was likely to cause inconsistency. The network models were capable of resolving the inconsistencies caused by injecting random noise into the input layers, i.e. they produced well-formed outputs despite receiving inconsistent inputs. Many of the outputs produced in

this way were errors. The errors produced broadly conformed to patterns of errors that the original subjects had produced. There were exceptions, however, and the simulations were not able to capture certain differences in error categories that depended on order of recall.

6.2 Successes

The work described here has been successful in developing experimental methodology and data modelling techniques. The methods developed have made it possible to investigate several important theoretical issues which increase our understanding of attribute binding. This section will first discuss the important methodological results and then discuss some of the theoretical advances that were made possible by the success of the methodology.

6.2.1 Methodological success

The methodology developed in this thesis can be broken down into the three main headings of *experimental methodology* — the development of the MIT, *statistical modelling* — the use of multiple regression on both the reading times and recall error data, and *PDP modelling*. This section discusses the development of these successful techniques. Section 6.2.2 discusses some of the theoretical insights that were gained by using them. Section 6.4 describes some of the areas opened up by the methods developed here.

Experimental methodology

The memory for individuals task has proved to be a versatile paradigm, providing both reading time and recall error data without any simple speed-accuracy trade-off. The descriptions used are contentful, requiring subjects to perform semantic processing in a way that must surely distinguish the task from any simple list

learning paradigm. Because there are always at least two individuals described and there is considerable overlap in vocabulary between descriptions, the task poses a difficult attribute binding problem. The level of difficulty of the task seems to be appropriate since, although most texts are recalled correctly, there are sufficient recall errors to allow statistical analysis. Although, like all such data, the reading times measured were fairly noisy, clearly interpretable patterns emerged. Thus, the paradigm has been a highly successful method of collecting data about attribute binding.

The reasons for the method's success are its ability to vary the structure of the information in the descriptions and the manageable variety of recall error patterns that occur. The manipulation most discussed in this thesis is that of matchtype, the pattern of matching and mismatching of vocabulary dimensions. The sensitivity of reading times to this manipulation supported the hypothesis that the semantic ordinal effect is due to the increasing difficulty of making appropriate associations as more is known about a particular individual. Since this work was done our research group has performed many experiments, employing broadly the same techniques, in which a variety of other structural manipulations of the description have been made. The most important of these experiments are described in Section 6.4.

Because the method yielded readily classifiable errors, it gave rise to a statistical model that could predict the frequency of an error category from the number of underlying features it maintains, the number it disrupts and their weightings or coefficients.

Statistical Analysis and Modelling

The statistical modelling techniques used on the reading time data were highly successful in partitioning the cognitive load according to simple structural aspects of the information in the experimental material. More sophisticated regression models have been built for more general descriptions in subsequent work by our

research group (see Section 6.4). Although the use of multiple regression in this way is not novel (e.g. see Kieras and Just 1984), the theoretical interpretations of these simple linear equations have allowed us to demonstrate how non-trivial a task attribute binding really is. The definition of plausible load variables which take effect locally or cumulatively and which can be tested against the data using multiple regression has been of crucial importance. It has facilitated the development of models of the reading time data of further experiments and extended our understanding of the dynamic cognitive load imposed by different text structures.

The regression modelling of the recall data was a surprisingly simple and successful method of accounting for the distribution of recall errors for the eight different matchtypes. The assumption that a greater degree of similarity between two stimuli leads to a greater degree of confusability and thus a greater probability of error is hardly new in cognitive psychology (e.g. Conrad 1964). However, the way in which multiple regression allowed us to extract a similarity metric in terms of independent and redundant features is novel. This simple modelling technique has allowed us to test whether our hypotheses concerning the redundancy and distribution of the underlying representation can account for the differences in frequencies between categories of recall error. The second recall model allowed us to simulate a component of the recall process within a PDP framework. The techniques used to build the second, untagged model have been used successfully for data in subsequent experiments (see Section 6.4).

PDP modelling

The PDP network was constructed to ensure that the theoretical interpretation of the recall model was sound and that a relatively simple mechanism was capable of performing inference from feature truth values to recalled vocabulary items. By and large, the back-propagation learning algorithm had few problems learning to perform this input-output mapping. What was perhaps more interesting was the performance of the trained networks when subjected to simple binary noise in their input vectors. Although the fit was not exact, the networks produced a

similar variety of error types to the original human data collected by the MIT.

In effect, the PDP framework allowed a statistical model of human data to specify a working model of a component of the recall process. A remarkably simple network used simple parallel constraint satisfaction inferences to map the underlying representation, as extracted by multiple regression, to correct and errorful recall. The least that such successful simulations do is demonstrate that a theoretical model works in practice. If the simulation is a clear embodiment of the theoretical ideas without too many additional parameters, then it demonstrates the adequacy of the theory in a particular instance, its explicit exposition and internal consistency.

The best simulations should facilitate the generation of new theoretical ideas and hypotheses. The PDP networks presented here focussed our minds on the importance of inconsistency resolution in the recall of the kind of representation that we claim underlies attribute binding (See Section 6.2.3). However, it is not yet clear which aspects of the performance of the model are important and which accidental. Thus, it is premature for new hypotheses generated by these simulations to be confidently considered. There are good reasons to suppose that the performance of the networks can be improved (see Section 6.3).

The General Modelling Approach

Throughout the thesis there has been an emphasis on *modelling*. The main statistical tool used, multiple regression, allowed exploratory model-fitting. A combination of theoretical justification and the coverage of variance in the data supported several variable definitions and types of models.

Conventional hypothesis testing methodology and the use of planned ANOVAs etc. has not been ignored (see pages 34, 54). All the experiments described have been designed so that certain hypotheses can be tested (e.g. the ruling out of articulatory rehearsal as an explanation for the semantic ordinal effect and the demonstration that non-binary vocabulary gives rise to a semantic ordinal effect).

The richness of the data from this paradigm is such, however, that a more powerful approach is required. Once a satisfactory model has been built it can be tested on further data. However, often the most interesting points are the differences between the models for different data. If the differences can be satisfactorily interpreted within the context of the general model then the modelling approach has borne fruit since it has provided a useful tool for suggesting new hypotheses. Examples of such comparisons in this work include the comparisons between the different reading time models in Chapter 2 and the comparisons made by Stenning, Patel and Levy (1987) between their models and the recall model presented here.

Parallel Distributed Processing provides a different kind of modelling framework. Here, the dangers lie in using an extremely powerful set of mechanisms simply to replicate a certain pattern of data without the exercise extending the explanation of a phenomenon. The work outlined in Chapter 5 places certain constraints on the network model. The model is specified by an already existing statistical analysis and the network simulation is kept as simple as possible to avoid losing sight of the main aim which was to explore possible recall mechanisms.

6.2.2 Theoretical advances

Apart from demonstrating that attribute binding is a real problem for human cognition by showing how manipulating its difficulty causes interpretable differences in reading times, perhaps the most important theoretical issues that were developed in the work described here were the influence of content and background knowledge and the redundancy in the underlying representation. We interpret differences in the difficulty of attribute binding as differences in the amount of work needed to search for associations based on previous knowledge that serve to bind the presented vocabulary items. We model these associations as redundant and fragmented features. Thus, the meaning (content) of the descriptions and the general (background) knowledge of the subjects are important aspects of the model. By assuming that attribute binding was achieved by a primitive contentless link, previous work on human knowledge representation had no chance

of dealing with these issues.

Content ...

Chapter 1 stressed the importance of content on the input, representation and recall of information in human memory. The work described in this thesis has concentrated almost entirely on structural aspects of the representation. This has been due to the practical issues underlying the design of the MIT. The descriptions used are designed so that their informational complexity (number of individuals, number of properties, matchtype etc.) can be easily manipulated. The ease of these manipulations and the sensitivity of reading times and recall errors to these structures has been what allowed empirical investigation at all. It is the theoretical interpretation of these essentially structural models that has made appeals to the use of background knowledge.

The interpretation put on the semantic ordinal effect and the regression models that successfully decomposed the structure of the descriptions into their load bearing components was one that involved content. The increase in reading times is interpreted as an increase in cognitive effort required to make the associations with the reader's general knowledge necessary to support the distinctions needed to allow a successful recall. Clearly, the difficulty of the task increases as the number of distinctions needed to support attribute binding increases and so mismatching dimensions increase reading times and matching dimensions impose a greater load once a mismatching dimension forces processing to focus on individuals rather than matching properties. Thus the investigational strategy of concentrating on the manipulation of structure has not diminished the crucial importance of background knowledge in the model. Although the model can account for variance in terms of structure, its interpretation demands the consideration of the meaning of the vocabulary items and the recruitment of the subject's general knowledge. What aspects of general knowledge are used is a matter of speculation. Our research group is making a start in the investigation of this area by testing the effects of the stereotypy of descriptions on subjects' performance on the MIT.

In a similar way to the reading times model, the recall model has a structural form but a content-based interpretation. A particular feature is true if all of its structurally or temporally defined vocabulary items are instantiated; otherwise the feature is false. The features are interpreted as contentful associations (i.e. associations linking the specific vocabulary items of the current description with previous general knowledge) that serve to make the necessary distinctions needed for correct recall. For example, the feature CA might represent an instantiation of the vocabulary items 'happy' and 'dentist' and the subject might remember a representation of a toothy smile which would be enough to recall that the dentist was happy when the recall menu was presented. This interpretation explains attribute binding in terms of constraint satisfaction between associations rather than an empty primitive link.

The PDP extension of the recall model has a localist structural representation of feature truth values as its input layer. Of all the components of the overall model, it makes the least appeal to a content-based approach. This is perhaps paradoxical considering the attractions of the framework discussed in Chapter 4 — many of which make it a suitable framework for modelling the influence of general knowledge. The use of the PDP framework here stressed simplicity. It was deemed better to begin with as restricted a model as possible rather than get lost in a sea of parameters. Possible ways of extending the PDP model in this direction are discussed in Section 6.3.

The theoretical approach to memory taken here appears to borrow methodology from the Ebbinghausian school of experimental psychology. The material appears similar to that used in word-list experiments (e.g Tulving 1983). However, the descriptions used are contentful and appeal naturally for the use of background knowledge. The experimental paradigm and analytical techniques have allowed a theoretical approach following Bartlett (1932). The success of the approach has been to uncover the details of the inferences required for attribute binding without ignoring the contentful nature of memory.

6.2.3 The Redundancy in the Representation

One of the most interesting successes of the recall error model is the redundancy in the representation that it specifies. This is caused by the overlapping of the independent fragmentary associations that support a subject's recall. A consequence of this redundancy is, as discussed in Chapter 5, that any disruption is likely to cause some inconsistency. This means that inconsistency resolution may be a truly pervasive component of cognition, not only necessary for the high level inconsistencies that cause belief revision but also necessary at the fundamental level of representing the attribution of properties to individuals. There has been much recent work in AI to produce systems capable of belief revision. It is to be hoped that developments of the kinds of PDP techniques presented here may prove useful for such problems (see Hinton et al. 1986, Shastri 1988).

The importance of the model is in its general characteristics rather than the precise combinations of vocabulary items that account for this particular data. It is clear that redundancy is involved and that there is evidence for a range of different sizes of fragments with no particular property in common. Models with similar but interpretatively different features have been found for different experiments (see Stenning, Patel and Levy 1987).

6.3 Possible Further Work

There are several weak points in the models described in this thesis. This section discusses how many of them might be strengthened by further work. Section 6.4 describes how the methodology described here has already allowed additional research by our research group.

There are several ways in which the PDP model of recall processes might be improved. One alteration to the network architecture that might improve the ability of the networks to generalise would be to restrict the connectivity of the input layer to the hidden layer (see Solla 1988). A sensible approach might be to

split the hidden layer into four groups of units, one for each vocabulary dimension and connect each input unit only to the dimensions that it instantiates. Another way to improve the generalisation performance would be to experiment with a greater number of hidden layers (see Wieland and Leighton 1987).

The present PDP simulation injects uniform noise into the input layer to induce errors on the output layer. The fit between model and data might be improved if the noise was shaped to reflect the mean accuracies of each feature for the errors that occur in the data. The model might also be improved if it took into account the coefficients for each feature variable in the regression model. The higher the coefficient the more disruption should be caused by its 'flipping'.

The PDP networks described in Chapter 5 use a localist representation for their input and output layers. This is convenient for their use in modelling the inferences required to produce well-formed recalls from consistent and inconsistent vectors of feature truth values. A more ambitious model might attempt to represent the features in terms of associations with some background knowledge. An interesting attempt to model subjects' judgements of the stereotypicality of combinations of the vocabulary items used within the MIT is reported in Nelson (1988). Such work might perhaps provide a basis for representing the features of the recall error regression model in a more distributed way in a PDP network.

Perhaps the most disappointing thing about the work described here is the failure to integrate fully the reading time and recall error models. They are not inconsistent with each other and the reading time model did give useful clues that were used in the development of the recall model. However, they remain distinct models and not different components of a single model. A possible way of rectifying this matter would be to focus on the constructive processes while bearing in mind the structure of the representation specified by the recall model.

One way to do this would be to reanalyse the reading time data using variables based on the instantiation of the features in the recall model. Another method would be to attempt to construct a process model of constructive processes within the PDP framework. This would hold the promise of integrating both constructive

and recall processes as well as a specification of a representation within a single model. There has been a great deal of recent work on the use of networks that can deal with temporal processing (e.g. Pineda 1987, Elman 1988, Servan-Schreiber et al. 1988). Hopefully some of this work would provide a basis for a suitable network model. Such a model might have a better chance of accounting for the recall order effects discussed in Chapter 5.

Naturally, it is hoped that this theoretical model has a wider scope than this particular experimental paradigm. Although they are relatively simple and always invite relational associations between the individuals, the texts used are meaningful natural language descriptions. Future work might include the use of even more realistic materials. Any loosening of the logical structure of the texts is likely to make reading times harder to model, but it is likely that recall error data will be analysable.

Multiple regression modelling of the different frequencies of recall errors is not the only method that might be used to model the recall data. Some preliminary work has shown that discriminant analysis shows some promise in modelling the distribution of representational features. Discriminant analysis produces a linear function of the independent variables for each category in the data, aiming to find the variables and functions that best discriminate between the categories. A successful discriminant analysis is one that misclassifies a small amount of data. So, in the context of modelling this data, multiple regression will pick out those features most important in predicting the frequency of the error categories, while discriminant analysis will pick those that best classify the different recall categories.

One way of using discriminant analysis here is to use a stepwise method to pick those features whose preservation or disruption best discriminates between the different categories of recall error. An initial analysis produced a fairly successful but highly unparsimonious model involving 80 variables. When the same 15 variables that were picked by the multiple regression procedure are used in a discriminant analysis the fit is not as good but still fairly respectable (83% of the

data being correctly classified), slightly better than a model produced from the first 15 variables in the first analysis. Encouragingly, the confusions made by these models are usually between closely related error categories.

Our expertise in this method is not yet far enough advanced to predict confidently whether it can improve the recall model. It shows promise as a way of modelling the origin of different error categories. Although perhaps of less theoretical importance than determining the structure of the underlying representation, the ability of a model to account for the relative frequencies of different recall errors is important. However, discriminant analysis does not seem appropriate to model this aspect of the behaviour of subjects and it has so far proved difficult to produce a parsimonious model using this method.

6.4 New Work

Since and during the time when the work described in this thesis was done, many other experiments have been performed by our research group. This section will briefly describe two areas where the development of the techniques described here made research possible.

The first of these areas has been the effect of different referential orderings of the sentences (i.e *modes*) in the texts of the MIT. Much of this work has been described in Stenning, Patel and Levy (1987) and Patel (forthcoming). The texts used in the work described here have been mostly Individual by Individual (IxI) or Predicate by Predicate (PxP). If these are the only text modes in an experiment, it is possible for a subject to discover the mode by the second sentence. When text modes are made more complex so that it is impossible to predict which individual the next sentence will describe, interesting reading time effects occur. While they still show the effects of matching and mismatching, the reading times of unpredictable texts are also increased when a switch of reference between two individuals is made. This increase is proportional to the number of properties known about the individual to which reference has switched. Based on recall error data, a

distinction was made between *primary* and *secondary* individuals. Regression analyses using this distinction in the definition of load variables accounted for significant reading time variance. Further research will determine whether these results add anything to the large literature on focus and foregrounding using other experimental paradigms.

Some of the experiments on unpredictable text modes lead to another area of work on the use of articulatory rehearsal. It appears that word length effects are far more important for unpredictable texts than for the predictable IxI and PxP texts described in Chapter 2. Although not completely accounting for the semantic ordinal effect, it is clear that articulatory rehearsal can be an important component for performance in the MIT. Evidence from the analysis of overt rehearsal protocols (Brown 1988, Nelson personal communication) suggests that rehearsal is most important for the maintenance of the secondary individual. Work is currently underway to study the effects of articulatory suppression (Baddeley 1986) on performance of the MIT. The experimental paradigm described in this thesis appears to be an excellent platform for the investigation of the workings of the different components of Baddeley's *Working Memory* framework. Reading times in the MIT are less noisy than those for more complex tasks and yet the task is much more realistic than word-list experiments. We are hopeful that the controlled but contentful nature of the MIT will allow us to ascertain the importance of different working memory components.

References

Anderson, J. R. and Bower, G. H. [1973] *Human Associative Memory.* Washington D.C.: Hemisphere.

Anderson, J. R. [1976] *Language, Memory and Thought.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Anderson, J. R. [1978] Arguments concerning representations for mental imagery. *Psychological Review*, **85**, 249–277.

Anderson, J. R. [1983] *The Architecture of Cognition.* Cambridge, Mass.: Harvard University Press.

Anderson, J. A. and Rosenfeld, E. (eds.) [1988] *Neurocomputing.* Cambridge, Mass.: MIT Press.

Baddeley, A. D. [1976] *The Psychology of Memory.* New York: Basic Books.

Bartlett, F. C. [1932] *Remembering: a study in experimental and social psychology.* Cambridge: Cambridge University Press.

Blake, A. [1983] The least disturbance principle and weak constraints. *Pattern Recognition Letters*, **1**, 393–399.

Brachman, R. J. and Levesque, H. J. (eds.) [1985] *Readings in Knowledge Representation.* Los Altos, Ca.: Morgan Kaufmann Publishers, Inc..

Bransford, J. D. and Johnson, M. [1972] Contextual prerequisites for understanding:some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, **11**, 717–726.

Bransford, J. D., Barclay, J. R. and Franks, J. J. [1972] Sentence memory: a constructive versus interpretive approach. *Cognitive Psychology*, **3**, 193–209.

Bransford, J. D. and Johnson, M. K. [1973] Considerations of some problems of comprehension. In Chase, W. G. (ed.) *Visual Information Processing*, pp389–392. New York: Academic Press.

Brown, J. [1958] Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, **10**, 12–21.

Carpenter, P. A. [1984] The Influence of Methodologies on Psycholinguistic Research. Chapter 1 in Kieras, D. E. and Just, M. A. (eds.) *New Methods in Reading Comprehension Research*, pp3-11. LEA.

Chan, L. and Fallside, F. [1987] An Adaptive Training Algorithm for Back Propagation Networks. Manuscript University of Cambridge, Department of Engineering.

Chomsky, N. [1965] *Aspects of the Theory of Syntax.* Cambridge, Mass.: MIT Press.

Clark, H. H. [1973] The language-as-a-fixed-effect fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, **12**, 335-359.

Collins, A. M. and Loftus, E. F. [1975] A Spreading Activation Theory of Semantic Processing. *Psychological Review*, **82**, 407-428.

Collins, A. M. and Quillian, M. R. [1969] Retrieval time from Semantic Memory. *Journal of Verbal Learning and Verbal Behaviour*, **8**, 240-247.

Conrad, C. [1972] Cognitive economy in semantic memory. *Journal of Experimental Psychology*, **92**, 149-154.

Conrad, R. [1964] Acoustic confusion in immediate memory. *British Journal of Psychology*, **55**, 75–84.

Cun, Y. L. [1985] Une procedure d'apprentissage pour reseau a seuil assmetrique. In *Proceedings of Cognitiva*, Paris, June, 1985, pp599–604.

Dixon, W. J. et al. (ed.) [1983] *BMDP Statistical Software.* University of California Press.

Draper, N. R. and Smith, H. [1966] *Applied Regression Analysis.* Chichester: John Wiley and Sons.

Draper, N. R. and Smith, H. [1981] *Applied Regression Analysis*, 2nd Edition.

Chichester: John Wiley and Sons.

Ebbinghaus, H. [1855] *Uber das Gedachtnis.* Leipzig: Duncker and Humblot.

Ebbinghaus, H. [1965] *Memory: A Contribution to Experimental Psychology.* New York: Dover.

Elman, J. L. [1988] Finding Structure in Time. CRL Technical Report No. 8801, Center for Research in Language, University of California, San Diego, April, 1988.

Fahlman, S. E. [1988] An Empirical Study of Learning Speed in Back-Propagation Networks. Technical Report No. CMU-CS-88- 162, Carnegie Mellon University, Pittsburgh, 1988.

Farhat, N. H., Psaltis, D., Prata, A. and Paek, E. [1985] Optical Implementation of the Hopfield Model. *Applied Optics*, **24**, 1469–1475.

Fodor, J. A. and Pylyshyn, Z. W. [1988] Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, **28**, 3–71.

Frauenfelder, U. H. and Tyler, L. K. (eds.) [1987] *Spoken Word Recognition.* London: MIT Press. Reprint from Cognition Vol 25 (1987).

Glass, A. L. and Holyoak, K. J. [1975] Alternative Conceptions of Semantic Memory. *COGNITION*, **3**, 313- 339.

Gemmell, M. D. [1988] Of Human Binding. Masters Thesis, Centre for Cognitive Science, University of Edinburgh.

Graesser, A. C. and Riha, J. R. [1984] An Application of Multiple Regression Techniques to Sentence Reading Times. Chapter 9 in Kieras, D. E. and Just, M. A. (eds.) *New Methods in Reading Comprehension Research*, pp183–218. Lawrence Erlbaum Associates.

Haberlandt, K. [1984] Components of Sentence and Word Reading Times. Chapter 10 in Kieras, D. E. and Just, M. A. (eds.) *New Methods in Reading Comprehension Research*, pp219–252. Lawrence Erlbaum Associates.

Hankamer, J. and Sag, I. [1976] Deep and Surface Anaphora. *Linguistic Inquiry*, **7**, 391–426.

Haviland, S. E. and Clark, H. H. [1974] What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal*

Behavior, **13**, 512–521.

Henderson, L. [1987] Word Recognition. A Tutorial Review. Chapter 8 in Coltheart, M. (ed.) *Attention and Performance XII: The Psychology of Reading*, pp171-200. LEA.

Hinton, G. E. and Anderson, J. A. (eds.) [1981] *Parallel Models of Associative Memory.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hinton, G. E. and Sejnowski, T. J. [1986] Learning and Relearning in Boltzmann Machines. Chapter 7 in *Parallel Distributed Processing*, Volume 1. Cambridge, Mass.: MIT Press.

Hinton, G. E. [1987] Representing part-whole hierarchies in Connectionist hierarchies. Unpublished manuscript.

Hirst, G. [1981] Anaphors in natural language understanding: a survey. pp128. New York: Springer-Verlag.

Hopfield, J. J. [1982] Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA*, **79**, 2554–2558.

James, W. [1892] *Psychology. Briefer Course.* London: Macmillan and Co.

Jarvella, R. J. [1971] Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, **10**, 409–416.

Johnson-Laird, P. N. [1983] *Mental Models.* Cambridge: Cambridge University Press.

Johnson-Laird, P., Herrmann, D. and Chaffin, R. [1984] Only connections: A critique of semantic networks. *Psychological Bulletin*, **2**, 292–315.

Johnson-Laird, P. N. and Bara, B. G. [1984] Syllogistic Inference. *Cognition*, **16**, 1–61.

Johnson-Laird, P. N. [1987] Connections and controversy. *Nature*, **330**, 12–13.

Johnson-Laird, P. N. [1988] *The Computer and the Mind: An Introduction to Cognitive Science.* London: Fontana Paperbacks.

Johnson-Laird, P. N. and Wason, P. C. [1977] *Thinking: Readings in Cognitive Science.* Cambridge: Cambridge University Press.

Jones, G. V. [1976] A fragmentation hypothesis of memory: cued recall of pictures and of sequential position. *Journal of Experimental Psychology*, **105**, 277–293.

Jones, G. V. [1978] Tests of a structural theory of the memory trace. *British Journal of Psychology*, **69**, 351–367.

Jones, G. [1984] Fragment and schema models for recall. *Memory and Cognition*, **12**, 250–263.

Kieras, D. E. [1981] Component processes in the comprehension of simple prose. *Journal of Verbal Learning and Verbal Behavior*, **20**, 1–23.

Kieras, D. E. and Just, M. A. (eds.) [1984] *New Methods in Reading Comprehension Research.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Kieras, D. E. [1984] A method for comparing a simulation to reading time data. Chapter 13 in Kieras, D. E. and Just, M. A. (eds.) *New Methods in Reading Comprehension Research*, pp299–326. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Kintsch, W. and Dijk, T. A. [1978] Towards a model of text comprehension and reproduction. *Psychological Review*, **85**, 363-394.

Knight, G. P. [1984] A survey of some important techniques and issues in multiple regression. Chapter 2 in Kieras, D. E. and Just, M. A. (eds.) *New Methods in Reading Comprehension Research*, pp13–30. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Landauer, T. K. and Meyer, D. E. [1972] Category size and semantic-memory retrieval. *Journal of Verbal Learning and Verbal Behavior*, **11**, 539-549.

Levy, J. [1988] Computers that learn to forget. *New Scientist*, No. 1625, 36–40.

Levy, J. and Stenning, K. [1988] A PDP Implementation of a Psychological Theory of Memory. In *First Annual INNS Meeting*, Boston, September, 1988, pp195. Extended abstracts only. Supplement to Neural Networks Volume 1.

Lowe, A. [1989] The Relative Contribution of Top-Down and Bottom-Up Information During Lexical Access. PhD Thesis, Department of AI and Centre for Cognitive Science, University of Edinburgh.

McClelland, J. L. and Rumelhart, D. E. [1985] Distributed memory and the representation of general and specific information. *Journal of Experimental*

Psychology: General, **114**, 159–188.

McClelland, J. L. and Kawamoto, A. H. [1986] Mechanisms of sentence processing. Chapter 19 in *Parallel Distributed Processing*, Volume 2. Cambridge, Mass.: MIT Press.

McClelland, J. and Rumelhart, D. [1986] A Distributed Model of Human Learning and Memory. Chapter 17 in *Parallel Distributed Processing*, Volume 2: *Explorations in the Microstructure of Cognition: Psychological and Biological Models.* Cambridge, Mass: MIT Press.

McClelland, J. L. and Rumelhart, D. E. (eds.) [1986] *Parallel Distributed Processing: Exploration in the Microstructures of Cognition*, Volume 2: *Psychological and Biological Models.* Cambridge, Mass.: MIT Press.

McCulloch, W. S. and Pitts, W. [1943] A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematics Biophysics*, **5**, 115–133.

McMillan, C. and Smolensky, P. [1988] Analyzing a Connectionist Model as a System of Soft Rules. Technical Report No. CU-CS-393-88, Department of Computer Science, University of Colorado, Boulder, March, 1988.

Miller, G. A. [1956] The magical number seven plus or minus two, or, some limits on our capacity for processing information. *Psychological Review*, **63**, 81–96.

Miller, G. A. [1962] Some psychological studies of grammar. *American Psychologist*, **17**, 748–762.

Miller, G. A. and Johnson-Laird, P. N. [1976] *Language and Perception.* Cambridge: Cambridge University Press.

Minsky, M. [1977] Frame-system theory. Chapter Wason in Johnson-laird, P. N. (ed.) *Thinking*, pp355-376. Cambridge: Cambridge University Press.

Minsky, M. and Papert, S. [1969] *Perceptrons: An Introduction to Computational Geometry.* Cambridge, Mass.: MIT Press.

Minsky, M. and Papert, S. [1988] *Perceptrons: An introduction to computational geometry*, 2. Cambridge, Mass.: MIT Press.

Morton, J. [1969] The Interaction of Information in Word Recognition. *Psychological Review*, **76**, 165-178.

Murray, A. F., Tarassenko, L. and Hamilton, A. [1988] Programmable Analog Pulse-Firing Neural Networks. In *NIPS '89 (Neural Information Processing Systems)*, 1988.Published as a book: Advances in Neural Information Processing Systems. 1989. Morgan Kaufmann. Palo Alto.

Nelson, S. [1988] A Simulation of Stereotypy in a Parallel Distributed Processing Framework. Masters Thesis, Centre for Cognitive Science, University of Edinburgh.

Norman, D. A. [1986] Reflections on Cognition and Parallel Distributed Processing. Chapter 26 in McClelland, J. L. and Rumelhart, D. E. (eds.) *Parallel Distributed Processing*, Volume 2: *Explorations in the Microstructure of Cognition*, pp531–546. London: MIT Press.

Parker, D. B. [1985] Learning-Logic. Technical No. TR-47, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, April, 1985.

Patel, M. [1989] Human text processing and models of knowledge representation. PhD Thesis, Centre for Cognitive Science, University of Edinburgh.

Peterson, L. R. and Peterson, M. [1959] Short-term retention of individual items. *Journal of Experimental Psychology*, **58**, 193–198.

Pineda, F. J. [1987] Recurrent Backpropagation. Technical Report No. S1A-63-87, Applied Physics Laboratory, The John Hopkins University, July, 1987.

Prince, A. and Pinker, S. [1988] On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition. *Cognition*, **28**.

Pylyshyn, Z. W. [1973] What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery. *Psychological Bulletin*, **80**, 1–24.

Quillian, M. R. [1968] Semantic Memory. In Minsky, M. (ed.) *Semantic Information Processing.* Cambridge, Mass.: MIT Press.

Reinhart, T. [1981] Definite NP anaphora and C-Command Domains. *Linguistic Inquiry*, **12**, 605–635.

Rosenblatt, F. [1958] The Perceptron, a probabilistic model for information storage and organization in the brain. *Psychological Review*, **62**, 386–408.

Rumelhart, D. E., Lindsay, P. H. and Norman, D. A. [1972] A process

model for long-term memory. In Tulving, E. and Donaldson, W. (eds.) *The Organisation of Memory*, pp197- 245. New York: Academic Press.

Rumelhart, D. E. and McClelland, J. L. (eds.) [1986] *Parallel Distributed Processing: Exploration in the Microstructures of Cognition*, Volume 1: *Foundations.* Cambridge, Mass.: MIT Press.

Rumelhart, D. E. and McClelland, J. L. [1986] On Learning the Past Tenses of English Verbs. Chapter 18 in *Parallel Distributed Processing*, Volume 2. Cambridge, Mass.: MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L. and Hinton, G. E. [1986] Schemata and sequential thought processes in PDP models. Chapter 14 in McClelland, J. L. and Rumelhart, D. E. (eds.) *Parallel Distributed Processing: explorations in the microstructure of cognition*, Volume 2: *Psychological and Biological Processes*, pp7–57. Cambridge, Mass.: MIT Press.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. [1986] Learning Internal Representations by Error Propagation. Chapter 8 in *Parallel Distributed Processing*, Volume 1. Cambridge, Mass.: MIT Press.

Sanford, A. J. and Garrod, S. C. [1981] *Understanding Written Language.* Chichester: John Wiley and Sons.

Sanford, A. J. [1985] *Cognition and cognitive psychology.* Lawrence Erlbaum.

Schank, R. C. [1982] *Dynamic Memory: A Theory of Learning in People and Computers.* Cambridge: Cambridge University Press.

Schank, R. C. and Abelson, R. [1977] *Scripts, Plans, Goals and Understanding.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Sejnowski, T. and Rosenberg, C. R. [1986] NETtalk, a parallel network that learns to read aloud. Technical Report No. JHU/EECS-86/01, Electrical engineering and computer science, John Hopkins University, Baltimore, 1986.

Servan-Schreiber, D., Cleeremans, A. and McClelland, J. L. [1988] Encoding Sequential Structure in Simple Recurrent Networks. Technical Report No. CMU-CS-88-183, Computer Science Department, Carnegie Mellon University, November, 1988.

Sivilotti, M. A., Mahowald, M. A. and Mead, C. A. [1988] Real-time visual computations using analog CMOS processing arrays. Chapter 43 in Anderson, J. A. and Rosenfeld, E. (eds.) *Neurocomputing: Foundations of Research*,

pp701–703. MIT Press. Originally published in Advanced Research in VLSI: Proceedings of the 1987 Stanford Conference. MIT Press.

Smolensky, P. [1987] On the Proper Treatment of Connectionism. Research Report No. CU-CS-377-87, Department of Computer Science and Institute of Cognitive Science, University of Colorado, Boulder, October, 1987.

Smolensky, P. [1988] The constituent structure of connectionist mental states: A Reply to Fodor and Pylyshyn. Technical Report No. CU-CS-394-88, Department of Computer Science, University of Colorado, Bboulder, March, 1988.

Solla, S. A. [1988] Learning Contiguity with Layered Neural Networks. In *First Meeting of the International Neural Networks Society*, Boston, September, 1988, pp223. Pergamon Press - supplement to Neural Networks Vol 1..

Stenning, K. [1975] Understanding english articles and quantifiers. PhD Thesis. University Microfilms.

Stenning, K. [1978] Anaphora as an approach to pragmatics. In Halle, M., Bresnan, J. and Miller, G. A. (eds.) *Linguistic Theory and Psychological Reality.* Cambridge, Mass.: MIT Press.

Stenning, K. [1980] On why making reference out of sense makes it so hard to make sense of reference. *Linguistics*, **18**, 619–633.

Stenning, K. [1986] On making models: a study of constructive memory. Chapter 7 in Myers, T., Brown, K. and McGonigle, B. (eds.) *Reasoning and Discourse Processes*, pp165–185. London: Academic Press.

Stenning, K. and Oaksford, M. R. [1989] Choosing Computational Architectures for Text Processing. Research Paper No. EUCCS/RP-28, Centre for Cognitive Science, University of Edinburgh, Edinburgh, April, 1989. To appear in Reilly, R. and Sharkey, N. (eds), Connectionist Approaches to Language, forthcoming.

Stenning, K., Shepherd, M. and Levy, J. [1987] On the construction of representations for individuals during text comprehension. Research Paper No. 9, Centre for Cognitive Science, University of Edinburgh, 1987.

Stenning, K., Patel, M. J. and Levy, J. [1987] The 'Binding Problem' in human memory: some effects of referential discontinuity on the construction of representations for individuals. Technical Report, Edinburgh University, 1987.

Stenning, K., Shepherd, M. and Levy, J. [1988] On the construction of

representations for individuals from descriptions in text. *Language and Cognitive Processes*, **2**, 129–164.

Stenning, K. and Levy, J. [1988] Knowledge-rich solutions to the binding problem: a simulation of some human computational mechanisms. *Knowledge Based Systems*, **1**, 143–152.

Touretzky, D. S. and Derthick, M. A. [1987] Symbol Structures in Connectionist Networks: Five Properties and Two Architectures. Research Report No. CMU-BOLTZ-26, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, 1987.

Tulving, E. [1983] *Elements of episodic memory.* Oxford: Oxford University Press.

Webber, B. L. [1978] A formal approach to discourse anaphora. PhD Thesis, Harvard University. Available as Report 3761, Bolt, Beranek and Newman Inc., May 1978.

Werbos, P. J. [1974] Beyond Regression: new tools for prediction and analysis in the behavioral sciences. PhD Thesis, Harvard University.

Werner, P. [1985] Comprehension of Definite and Indefinite Text and Visual Sequences: A Study in Constructive Memory. Masters Thesis.

Wieland, A. and Leighton, R. [1987] Geometric analysis of neural network capabilities. In *IEEE international conference on neural networks*, 1987.

Appendix A

Experimental Materials

A.1 The Antonymy Experiment

A.1.1 Binary Vocabulary

Choose one pair from each cohort:

A:

nurse	priest
judge	monk
vet	chef
doctor	vicar
teacher	bishop
dentist	baker

B:

French	Greek
Welsh	Swiss
Dutch	Czech
German	Spanish
Chinese	Polish
Swedish	Russian

C:

young	old
fat	thin
rich	poor
clever	stupid
hungry	thirsty
greedy	clumsy

D:

tall	short
strong	weak
sane	mad
friendly	hostile
happy	gloomy
daring	timid

A.1.2 Non-binary Vocabulary

Choose one from each list for each cohort:

A:

nurse	priest	doctor	vicar	vet	chef
judge	monk	dentist	baker	teacher	bishop

B:

French	Greek	German	Spanish	Welsh	Swiss
Dutch	Czech	Chinese	Polish	Swedish	Russian

C:

young	old	clever	stupid	fat	thin
rich	poor	hungry	thirsty	greedy	clumsy

D:

tall	short	sane	mad	friendly	hostile
strong	weak	happy	gloomy	daring	timid

A.2 The Replication Experiment

A.2.1 People Vocabulary Set

This was identical to the Binary Vocabulary above.

A.2.2 Object Vocabulary

A:

circle	square
triangle	oval
rectangle	ellipse
beam	block
cylinder	pyramid
disc	cube

B:

black	white
silver	gold
red	green
shiny	dull
yellow	blue
rough	smooth

C:

old	new
hard	soft
rough	smooth
wet	dry
hot	cold
light	heavy

D:

thick	thin
deep	shallow
hollow	solid
large	small
long	short
wide	narrow

Appendix B

Computer Software Development

A large amount of effort was expended in the development of several different software packages.

The hardware platform used for most of the recall error categorisation and PDP simulation was a SUN 3/160 workstation equipped with a 68881 floating point co-processor. The speed of the machine along with its UNIX operating system and flexible windowing interface combined to make an effective environment.

The recall error categorisation and feature scoring was performed using a package written in PROLOG. This language was chosen because of its flexible data structure manipulation abilities and its declarative rule-based nature. This allowed new definitions of error categories or feature definitions to be simply added to the database without major modification of the program. The package provided flexible primitives with which to describe the structure of the descriptions and the differences between a stimulus and a recall. The version of PROLOG that was used allowed programs to be compiled to speed up the program runs.

The PDP simulation package was written in the C programming language. This language is reasonably convenient for numerical computation and allowed an easy interface between the program and the UNIX operating stsyem and windowing interface. The program allowed the use of an arbitrary network structure and training set. All conventional parameters were easily altered and the error curves and weight arrays could be recorded at any point. Several extra capabilities were built in to specifically allow the training and simulation runs described in the thesis to be performed conveniently. The program was designed to interface with a graphics tool.

The graphics tool was developed with Dr. Andrew Zisserman of Oxford University.

It used the windowing environment provided by the workstation to full effect. The tool allowed the state of every node to be examined for any learning trial. It also allowed the input nodes to be 'flipped' and the resulting disrupted input to be fed into the network. The use of graphics allowed easy 'debugging' of networks during training as well as allowing a closer look at a mature network.

Appendix C

Published Journal Papers

LANGUAGE AND COGNITIVE PROCESSES, 1988, *3* (2) 129–164

On the Construction of Representations for Individuals from Descriptions in Text

Keith Stenning, Martin Shepherd, and Joe Levy

Centre for Cognitive Science and Department of Psychology, University of Edinburgh, Edinburgh, U.K.

Subjects read texts describing pairs of individuals sentence by sentence in a self-paced reading time task and then answered questions about the individuals and recalled them. The task is designed to explore how people represent the binding of attributes to individuals.

The reading time data show that the construction of representations is organised around what is known about the currently referenced individual. The more that is known the more slowly subjects read. This slowing is not due to articulatory rehearsal. A regression model of reading times describes the partitioning of working memory resources across the semantic structures being processed.

The construction processes yield redundant representations consisting of sets of feature values encoding aspects of the information in the text and contributing independently to memory performance. Modelling in terms of feature representations enables prediction of the patterns of error in recall.

The reading time and recall error models are interpreted as showing how processes of recruiting associations occupy increasing time as more properties are known of an individual.

INTRODUCTION

Suppose that we learn that there is a bishop. Later we learn that he is Polish, that he is tall, and that he is hungry. We then know that there is a tall hungry Polish bishop. How do we encode these successive pieces of information? How do we access this information when we recall the person? When we learn of a Swiss dentist, how do we remember that it was the bishop who was Polish? All these questions lead to one central

Requests for reprints should be addressed to Keith Stenning, Centre for Cognitive Science, 2 Buccleuch Place, Edinburgh, U.K.

This research was supported by grant no. D/10138 from SERC/Alvey and by an ESRC linked studentship to Joe Levy.

question: How are several properties represented as belonging to one and the same individual?

We will call this the "attribute binding problem" or simply the binding problem for short. It is an extremely general problem in knowledge representation and, therefore, memory. It is closely related to the anaphor resolution problem which has received much more attention (see Hirst, 1981, for a review). Anaphor resolution is the process of selecting one of several possible antecedents for an anaphor (such as a definite pronoun) while understanding a text. Binding is the process of then representing the results of such assignments. Binding has received less attention, probably because it is easy to achieve by structural means in conventional computer architectures. However, for human beings, binding can present great difficulties and many classical results demonstrate that it is solved by human beings by knowledge-rich means (e.g. Miller, 1959). The probability of remembering that X has property F, whereas Y has property G, is a function of the *content* of X, Y, F, and G.

The importance of resolving the mapping of properties on to individuals for comprehension, reasoning, and memory has been stressed, for example, by Stenning (1978; 1986) and Johnson-Laird (1983). Texts introduce individuals and then progressively map properties on to them. The conventions of expository text normally ensure that speakers provide information in a manner in which a definite mapping can be constructed at all points in a text. In a logician's sense, such texts determine unique models. Reasoning problems which demand the consideration of several distinct models are much harder to solve than ones whose solution only requires one model to be considered. But we know little about how such mappings/models are represented in memory. To claim that there are mental models is to claim that these mappings have a special status in human cognitive processes but not to explain how they are represented or manipulated. Our aim is to investigate this important component of text comprehension and memory. Our focus on these referential aspects of text lead us to study texts of stark simplicity, stripped of all problems of anaphor resolution, and of all intensional complexities. What remains is not the whole of the problem of text comprehension, but it is a most important and somewhat neglected component.

The resulting structures, descriptions of a small set of individuals in terms of a few attributes, are intermediate between natural text and the unstructured lists used in much of the episodic memory literature (e.g. Tulving, 1983). Consider an example:

> There is a bishop. The bishop is Polish. The bishop is tall. The bishop is sad. There is a dentist. The dentist is Swiss. The dentist is tall. The dentist is happy.

This text is comparable to an eight-item list of content words, but the addition of referential structure is what poses the binding problem and leads to the different memory phenomena to be described here. The vast majority of the episodic memory literature avoids explicit study of the binding problem. Studies of unstructured or even of categorised lists involve binding only in that item occurrence must be bound to list identity. But since retrieval generally follows immediately on each list, free recall emphasises the problem of item retrieval, and the identity criteria for lists are implicit contextual features, there is little opportunity for explicitly studying this binding. Some work has acknowledged this lack of structure by studying memory for texts where several properties are attributed to several individuals. Anderson and Bower (1973) and Anderson (1983) have developed semantic network theories of the representations involved in text processing, using relatively naturalistic texts. These clearly involve structure, but present only slight binding problems. In their materials, few combinations of properties with individuals are plausible on general knowledge grounds. When the hippy kisses the debutante who sits on the bench we do not have to remember that the debutante didn't sit on the hippy or kiss the bench because we "knew" these things already. Most of the binding is already done in general memory. The theories arising from this work use primitive representational elements ("links") to solve the binding problem in a content-independent manner, as do a range of theories of knowledge representation which are quite distinct on other dimensions (e.g. Johnson-Laird, 1983; Rumelhart, Lindsay, & Norman, 1972; Schank & Abelson, 1977). Much attention is given to "fan-effects" of the numbers of properties linked to a node, but binding is not seen as a problem of much interest in itself.

The work of Jones (1976) does raise the issue as to whether primitively unanalysable links in wholistic network structures are appropriate models of the binding that is demanded by Anderson's material. Jones shows convincingly that representations consisting of independent "fragments" of experiences predict cued recall data better than network models. Jones' material consists of series of pictures of objects which are of several types, have colours, and are in spatial locations and temporal positions. Subjects perform cued recall after a sequence of such pictures. The pictures present all attributes simultaneously so one cannot study the progressive construction of representations of binding. Only one individual ever occurs in a picture, and the materials are designed so that there is minimum overlap between the properties of any two individuals. So this paradigm does not place as strong an emphasis on binding as might at first appear. The data bear this out in that few intrusion errors of one individual's properties into recall of another individual occur. But Jones is almost alone in making the important point that different systems for representing individuals exist

and require exploration. Our own conclusions will be compared with those of Jones in our discussion.

We have developed a novel Memory for Individuals Task (MIT) specifically to explore human memory solutions to the binding problem as it arises in texts which progressively specify individuals (see Stenning, 1986, for a preliminary exploratory experiment). By simultaneously presenting information about several individuals who may share several properties, we maximise the binding problem. Like Jones' material we use structured dimensions (profession, nationality, temperament, stature for people, shape, colour, texture, size for objects) each represented by a small number of values. Pairs of individuals are constructed by selecting values from each of these dimensions. Because all selections make coherent descriptions of individuals, subjects must distinguish between large numbers of possible combinations in memory. Because some individuals' properties overlap extensively, interference is severe. When a subject is faced with a recall menu after one such pair of individuals has been described, there are 136 unordered pairs of individuals available as responses. On the other hand, the material is rich in general knowledge associations, and this allows subjects to perform this task rather accurately. In the long term, we are concerned to understand how this content determines binding, but in this paper our analyses will be at a structural level. We focus on the structure of two individuals described on four dimensions, and will not here analyse differences in content within a dimension across paragraphs. It is nevertheless an important constraint on even structural analyses that they be consistent with accounts of the content dependence of binding in human memory.

Because text presents the properties of individuals progressively and several individuals' descriptions may be intertwined, the process of the incremental construction of representations is central to an understanding of both structures and processes. Questions arise about the partitioning of working memory resources between individuals, and about the interface between working memory and long-term memory. Linguistic studies of anaphor resolution show that not all antecedents are equally available for reference at all points in a text (Sanford & Garrod, 1981; Stenning, 1978) but little detailed information is available about how much of what representations are available in what sorts of memory as a reader processes text.

In order to construct a theory of the processes and structures involved in the extensional aspects of text comprehension, a methodology is required which produces rich data and constructs models with internal structure. Simple binary hypothesis testing is a weak approach to complex systems. Even the earliest pilot data from the MIT showed radical differences from that obtained in list learning experiments: the usual serial position effects

do not occur. Proactive interference is weak or non-existent. The resulting memory is well integrated and quite durable. One approach often advocated in such circumstances is through AI modelling (Newell, 1973). However, in the area of episodic memory such methodology has a crucial weakness, i.e. it is too easy to get a conventional computer to perform episodic memory tasks. Binding in particular is a trivial task in such architectures. To get the computer to exhibit the same weaknesses as the human subject must be done in a principled way if it is to be of much interest. Even where modelling approaches have been adopted, they have focussed on details of implementation such as "fan effects" without asking more general questions about a wider range of representational questions. This approach has tended to lead to scepticism about the possibility of resolving representational issues (e.g. Anderson, 1978).

We choose instead to use a more data-driven statistical modelling approach through multiple regression (see Kieras, 1981), and to investigate several sources of data about the same mental structures and processes in parallel. Regression forces the equation builder to find factors which behave independently of each other and focusses on the modularity of processes and structures. Through residuals analysis it reveals where the data diverge from the modularity formulated in an equation. It also measures the extent to which an account of the phenomena has been achieved. By modelling both reading times and error patterns resulting from the same memory experiences we seek to constrain accounts of both working memory and long-term memory representations and processes associated with solutions to the binding problem. This methodology is above all exploratory, seeking to construct models of complex processes from as broad an evidence base as possible. From these regression models it is possible to construct PDP models (Stenning & Levy, 1988) in a computational architecture in which memory/inference limitations arise as an organic part of the computational process rather than as add-on degradations of performace (see McClelland & Rumelhart, 1986, for a general introduction to PDP systems).

What general questions about the representation of binding should an exploratory methodology explore? A general cluster of related issues about memory representations is whether they are redundant, whether they are wholistic and, if they have internal articulation, how are they structured and which parts are dependent and which independent of each other. These questions are particularly evident when considering representations of individuals constructed from separate elements. Is the representation structured along the same lines as the information supplied in progressive description? Semantic network theories assume, like the logical theories which they mimic, that the answer is "yes": A node is constructed to represent the newly introduced bishop, and subsequent

anaphoric references to the bishop serve to set up links to the new properties those references attribute to him. The representation is articulated in the same way as the description. This obviously does not have to be. Jones' fragments do not mimic this structure: Individuals' properties are linked together into fragments according to some probability of association between the dimensions (so colour is often linked to object-type to form a fragment, but not often to place). But the fragments of various sizes are themselves just that, fragments independent of all other fragments. In our material, several individuals appearing in the same context may overlap in their attributes so that pairs of their properties are not in general sufficient to discriminate them. Do people still employ fragmented representations of them, or do they resort to something more structured? We present here new methods of analysing error frequency data which is designed to explore these questions.

In analysing the reading times from the MIT to reveal the processes and structures of working memory, we build on Stenning (1986). Two phenomena are revealed by reading time measures of the self-paced, sentence by sentence reading of these texts. Reading time is almost wholly determined by the number of properties known of the individual referred to by the current sentence, and is littled affected by the the number of properties known of other individuals. Reading time increases as more properties are known of the individual referenced by the current sentence. We refer to these phenomena jointly as the semantic ordinal effect (SOE): "semantic" because it is *reference* which determines processing, and "ordinal" because it is the position in the sequence of attributions to an individual which determines time spent. These effects can be interpreted because two different orderings of presentation of the attributes, or modes, were used: individual by individual (I × I) and property by property (P × P). The former has been illustrated above, and the latter is given here:

> There is a bishop. There is a dentist. The bishop is Polish. The dentist is Swiss. The bishop is tall. The dentist is tall. The bishop is sad. The dentist is happy.

In P × P texts, reference switches between a pair of individuals on every sentence, yet this has little effect on reading time, which is still determined by the number of attributes known of the referenced individual.

The SOE naturally suggests questions about the role of rehearsal in this task. If subjects rehearse the properties of the referenced individual, more properties will take more time in proportion to the number of syllables and this might account for the SOE (Baddeley, Thomson, & Buchanan, 1975). There are several distinct questions here. Does rehearsal take place? Is it the basis for memory during the reading of the text? Does it account for the

SOE? The answer to the first is almost certainly "yes", subjects can quite often be heard rehearsing, and presumably many rehearse covertly. The answer to the second is that syllabic rehearsal may play some role in memory during reading, but the modes' similarity of reading time pattern suggests that other forms of memory must operate as well. In a P × P text, a rehearsal loop basis for memory is inadequate. The rehearsal of several properties of one individual would intervene between two cycles of rehearsal of the properties of the other individual, and this would be contrary to everything that is known of articulatory rehearsal as a basis for item memory. The question of what role rehearsal plays is taken up in Stenning, Patel, and Levy (1987). The answer to the third question, whether syllabic rehearsal accounts for the SOE, will be taken up here. Fortunately, syllabic rehearsal has a diagnostic footprint, i.e. the word length effect. In the present experiment we included texts made up of one- and two-syllable content words to measure the contribution of syllabic length to reading time.

We interpret the semantic ordinal effect as being due to increases in semantic integration processing required by increasing numbers of properties. Having established that these processes are organised around the properties of the currently referenced individual, we ask general questions about them first, in order to progressively specify what is consuming processing time. One issue that arises is the balance between maintenance and constructive processes. Information is taken in a superficial form, and a semantic representation is constructed. The superficial information must be maintained while the construction of semantic representations takes place. Consider again I × I texts like the example cited above, in which the subject reads a string of consecutive statements about an individual. The number of such statements before a change of individual is predictable in these experiments. If readers pursued a strategy of maintenance of all information in a superficial form until the last statement about the individual, the time to construct a representation for the last individual would be reflected in the final sentence reading time. It should be roughly equivalent to the time that it would take to process the information presented all at one exposure. On the other hand, if readers do as much processing as possible before taking in the next sentence, one would expect the time taken to process all the information at once to be substantially more than the time to process the fourth attribution alone. Comparison of the two situations affords the opportunity to find out where, between these two extremes, subjects actually operate. Accordingly, in addition to I × I and P × P modes, a third *multiple attribution* mode (MA) was included in which all four properties of an individual were presented in one sentence (e.g. "There is a tall sad Polish bishop").

Finally, subjects commonly report an intuitively salient strategy of

focussing on the matching and mismatching between the individuals on the various property dimensions. One of the aims of regression modelling of the reading time data is to see whether it can reveal a consistent account of this strategy. Although it constitutes an "artificial strategy" that is not applicable to texts in which these structural relations are replaced by contentful relations, matching and mismatching relations offer distinct methodological opportunities. Because they are repeated in each text and for each dimension, they can be analysed in ways unavailable for more varied material.

METHOD

Design

Subjects read texts consisting of up to eight simple declarative sentences. The texts described two objects (Object vocabulary set) or two people (People vocabulary set). There were three modes of text presentation—*Individual by Individual* (I × I), *Predicate by Predicate* (P × P), and *Multiple Attribution* (MA). Examples for the people vocabulary set include:

(I × I) There is a nurse. The nurse is French. The nurse is young. The nurse is strong. There is a chef. The chef is Greek. The chef is old. The chef is strong.
(P × P) There is a nurse. There is a chef. The nurse is French. The chef is Greek. The nurse is young. The chef is old. The nurse is strong. The chef is strong.
(MA) There is a strong young French nurse. There is a strong old Greek chef.

Note that in the MA mode the order of properties corresponds to the most natural adjective ordering, whereas in the I × I and P × P texts properties appear in reverse order.

Individuals were always mismatched on the first property (e.g. nurse versus chef in the above examples). Apart from this initial mismatch, individuals were matched on 0, 1, 2, or 3 properties equally often. When individuals matched on 1 or 2 properties, the position of the matching property in the text was random.

For the People vocabulary set texts consisted entirely of one-syllable words or entirely of two-syllable words.

After reading each paragraph, subjects answered two questions. Questions took the form of a noun paired with an adjective (e.g. "Is there a Greek nurse?"). The noun corresponded equally often to the first and

second individuals, the answer was equally often "yes" or "no", and the questioned dimension was matched and mismatched equally often across the individuals. Within these constraints, the property was selected at random from the model.

After answering the questions, subjects recalled the individuals from the paragraph they had just read.

Vocabulary

For each vocabulary set, properties were grouped into four "dimensions", corresponding roughly to shape, colour, texture, and size for the Object vocabulary set, and occupation, nationality, physical character, and temperament for the People vocabulary set. The dimensions also generally determine a natural (or at least a plausible) adjective order (dimension D to dimension A). Each dimension contained 12 pairs of contrasted adjectives or nouns, the adjectives being mostly antonymous pairs. The materials for the People vocabulary set were divided into one-syllable and two-syllable groups. Our semantic requirements make it too difficult to meet this syllable constraint in the object vocabulary. The complete set of materials is listed in Appendix 1.

Subjects

A total of 24 psychology students participated as part of a course requirement.

Procedure

Subjects were tested in groups of 12, using a network of BBC model B microcomputers. Each subject performed in one morning session on one vocabulary set, and in one afternoon session on the other vocabulary set. Subjects were given verbal instructions which explained the procedure. It was emphasised that subjects should take as much time to read each sentence as they felt was necessary to give correct recall, that they should answer the questions as quickly as possible consistent with accuracy, and that in the recall phase they should take plenty of time and try to be as accurate as possible.

Reading. Each session consisted of 48 texts. Each text was preceded by a "setting" which displayed the dimensions on which the individuals would be classified. For example, a text in the Objects vocabulary set might have this setting:

old/young—strong/weak—Greek/French—nurse/chef

The setting remained visible until the subject pressed the space bar, when it was replaced by the first sentence of the text (e.g. "There is a chef"). Reading of the text was self-paced, with the subject pressing the space bar to obtain each sentence. Each sentence replaced the previous one, so that the previously read text was not available to the subject.

Questions. At the end of the text, a warning message appeared, instructing the subject to place a forefinger of each hand on the response keys (A for yes, N for no) in readiness to answer the question. After a delay of 1.2 sec, the question (e.g. "Is there a strong chef?") appeared, and remained visible until the subject responded. No feedback was given for responses to questions.

Recall. Then a recall menu was presented, and subjects were prompted for the complete recall of each individual. Instructions (presented with the menu) emphasised that they could recall the individuals in any order they wished, but that the order of recall of properties within an individual should follow the order given in the menu (ending with the nominal property, e.g. nurse or square).

Subjects typed in their recall using the number codes given in the menu. Some correction (within an individual) was possible. Recall was followed by a display of the individuals that were actually presented in the text, in single sentence form (e.g. "There was a strong young French nurse and a strong old Greek chef"). No other feedback about the actual accuracy of subjects' recall was given. Subjects pressed the RETURN key to initiate the next text, which followed after a delay of 1.6 sec.

RESULTS AND DISCUSSION

Subjects who failed to attend both sessions were excluded, leaving 17 subjects in the following analyses. Because both this sort of data and the approach to analysis are novel, some general characteristics of performance will be presented before the detailed development of an overall model is undertaken. As a preliminary, the simultaneous use of reaction time and error data requires some comment.

In simple choice tasks, speed/accuracy trade-off complicates the simultaneous analysis of reaction times and errors. In a task such as the present one, the complexity of the responses, and the fact that the timed response is not the choice response, changes the situation. Subjects are being asked to spend as much, and only as much time as they need, reading a sentence in order to ensure accurate recall. For investigation of the representations underlying performance, the chief source of data is correlations between errors in different parts of the resulting structures. Even if there was a

point by point correspondence between reading time for a piece of information, and a likelihood of error on recalling that piece of information, this would not obstruct our chief interest which lies in the correlation of errors across different parts of the structure. Such point by point correspondence is, in fact, weak or non-existent (see below).

Recall accuracy was negatively correlated with total paragraph reading time—$R = -0.25$, $P < 0.0001$; mean reading time (sec) = 2.13 + (recall score × −0.167)—and subjects read paragraphs they made errors on more quickly than those they recalled correctly. For the I × I and P × P modes, this effect can be analysed by sentence position, and it is then possible to ask whether it is a local effect whereby spending less time on a sentence selectively prejudices the accurate recall of its content, or whether it is a global effect whereby reading a sentence quickly prejudices the accuracy of recall of the whole paragraph.

At each point, texts that were correctly recalled with regard to that sentence's information were read more slowly than texts recalled incorrectly, though the size of the effect was small for the two indefinite introducing sentences. However, in all but one sentence position, these appear to be global effects. Only at sentence four of individual one was there a pronounced local difference of reading time for that sentence between texts where an error was made on that property and ones where it was not (0.99 sec). This local effect cannot be accounted for by globally fast reading times (the texts containing errors were on average read only 0.21 sec faster per sentence than completely correctly recalled texts). We take these facts to justify our simultaneous use of self-paced reading time and accuracy of recall measures.

In what follows, a recall individual is an individual specified in a subject's recall protocol, and a stimulus individual is an individual specified by a stimulus paragraph. The subjects' recall was scored as follows: The best fitting assignment of recall individuals to stimulus individuals was chosen unless the two possible assignments were equally bad. In this latter case, the recall was treated as being in the same order as that of presentation. This decision is based on the observation that there was a strong tendency to recall the individuals in the order in which they were presented. Unclear cases are therefore resolved in favour of this observation from the clear cases.

An expositional decision has been taken to organise the data around the recall individuals rather than the stimulus individuals. This decision rests on the observation of patterns of errors in the data. From here on, "individual" will refer to recall individuals unless otherwise stated. The introducing property (shape or profession) will be lettered A, the next nearest in adjective order (colour or nationality) B, the next (texture or temperament) C, and the earliest in adjective order (size or stature) D.

In general, the recall was quite accurate. A total of 71% of all paragraphs were recalled entirely correctly. The mean number of properties wrong per paragraph was 0.55 (S.D. = 1.0). An analysis of variance was carried out with the factors: vocabulary (people/objects), practice (first half/second half), mode (I × I/P × P/MA), individual (first recalled/second recalled), property (A–D), and subjects, with the mean accuracy of recall as the independent variable. There were significant main effects of individual ($F = 108.31$, $P < 0.0001$; means first/second recalled individuals: 0.052/0.085), property ($F = 3.9$, $P = 0023$; means for properties A/B/C/D: 0.060/0.062/0.075/0.079), and of half (of the experiment) ($F = 8.29$, $P = 0.024$; means first half/second half: 0.052/0.085). There was a significant interaction between individual and property ($F = 3.90$, $P = 0.023$). The means for the four properties of the first individual were 0.056/0.040/0.052/0.062, and for the four properties of the second were 0.063/0.084/0.098/0.096. There were no other significant effects. Since there is no main effect or interaction with vocabulary or mode, nor any interaction between practice and any other factors, the following analyses pool across these factors and concentrate on the effects and interactions of individuals and properties. Investigation of interference effects within the two halves of the experiment similarly revealed little or no proactive interference.

INTERDEPENDENCIES BETWEEN ERRORS

Because this task is a new one we give descriptive analyses of errors within property dimensions and within individuals in Appendix 2. The choice of modelling approach to be described now emerged from these analyses. To summarise the findings, the data reflect strong correlations between the recall of some parts of these structures, but no correlations between other parts. There are strong correlations between errors on the two individuals within the same property dimension. There are strong correlations between errors on some of the properties within an individual. The least nominal property (property D) is the most often involved in these latter correlations, and the introducing property (property A), the only one actually presented along with the others, is least often implicated in these multiple errors. An explanation of this observation must be one of the benchmarks for any model of recall performance in this task.

We now turn to the problem of how errors are distributed within the whole structure of the pairs of individuals. Because there is such a large number of possible responses to each stimulus (136), classification is imperative. The adopted classification is developed in Appendix 2. In choosing a classification of error types we pay attention both to their observed frequency and their theoretical interest. For example, there are four types of double errors: *individual polarity, property polarity, double*

TABLE 1
Some Examples of Recall Errors

Response Type	*Response*							
Correct	tall	happy	Polish	bishop	short	happy	Swiss	dentist
Single	**short**	happy	Polish	bishop	short	happy	Swiss	dentist
Individual polarity	**short**	happy	Polish	bishop	**tall**	happy	Swiss	dentist
Property polarity	tall	**sad**	Polish	bishop	short	**sad**	Swiss	dentist
Double complementary	tall	**sad**	Polish	bishop	short	happy	**Polish**	dentist
Double homogeneous	**short**	happy	**Swiss**	bishop	short	happy	Swiss	dentist

complementary, and *double homogeneous*. An individual polarity error is a double error caused by making the wrong assignments for a mismatched dimension. A property polarity error is a double error caused by recalling the wrong vocabulary item for a matched dimension. A double complementary error is a double error caused by one error on a matched dimension and one on a mismatched dimension. A double homogeneous error is caused by errors on two matched or two mismatched dimensions. Table 1 shows some examples of these error types for a single presented

TABLE 2
Observed Probabilities of Occurrence and Sizes of Response Categories

Abbreviation	*Response Type*	*Observed*	*Size of Category*
corr	Correct	0.707	0.0007
misc	Miscellaneous	0.014	0.569
sg1+	Single error on I-1 matched	0.018	0.009
sg1−	Single error on I-1 mismatched	0.025	0.015
sg2+	Single error on I-2 matched	0.024	0.009
sg2−	Single error on I-2 mismatched	0.044	0.015
ipol	Individual polarity error	0.066	0.015
is1+	Individual polarity with "sg1+"	0.005	0.016
is1−	Individual polarity with "sg1−"	0.004	0.023
is2+	Individual polarity with "sg2+"	0.012	0.016
is2−	Individual polarity with "sg2−"	0.008	0.023
2cs1	Double complementary both on I-1	0.008	0.019
2cs2	Double complementary both on I-2	0.014	0.019
2cdf	Double complementary on I-1 and I-2	0.007	0.032
dhs1	Double homogeneous on I-1	0.004	0.019
dhs2	Double homogeneous on I-2	0.007	0.019
dhdf	Double homogeneous on I-1 and I-2	0.002	0.037
ppol	Property polarity error	0.012	0.009
pp+s	Property polarity with single	0.005	0.055
mirr	Mirror image matching structure	0.008	0.049

paragraph. Table 2 shows the final selection of 20 error types used in subsequent analysis, the proportion of responses observed of each type and the "size" of each category. The latter is the proportion of the 136 possible responses that fall into each category.

Examination of Table 2 reveals several more phenomena in need of explanation. Individual polarity errors are much commoner than property polarity errors. Individual polarity errors frequently occur in combination with single errors on other dimensions despite the fact that there are then three properties simultaneously wrong. Homogeneous double errors are much less common than complementary double errors despite the fact that there are roughly equal opportunities for the two types and they are both double errors.

BUILDING AN INTEGRATED MODEL OF RECALL PERFORMANCE

The descriptive analyses of the error data show that some parts of these structures share common fate in memory whereas others are independent of each other, and suggest that there is redundancy in the representations. For example, matching information is represented. A modelling framework is required which will characterise patterns of dependency between the parts of these structures. Jones' fragmentation models do this for the single individuals with four properties which he used. Even though our pairs of individuals with four properties are still very simple, Jones' method of modelling is not manageable for this material. The number of types of cueing required to estimate the efficacy as a cue of each of the different types of fragment would be too large. The presence of matching information in memory also suggests that there is redundancy in memory, and the fragmentation framework does not approach the problem of redundancy.

We have developed the following framework for analysing the "features" present in memory in this task. We use the term "features" for functions which take values for every pair of individuals presented or recalled. So, for example, one possible feature is "Nationality and temperament of the first individual" and it takes such values as "happy and Polish", "sad and Polish", "sad and Swiss", etc. Another example feature is "matching status of the nationality dimension" which would take values "matched" or "mismatched". Neither of these features are ever explicitly presented in the texts, though they can be inferred from what *is* presented. For any feature we might define, and any observed presentation/response sequence in the data, the feature either keeps the same value (it is preserved in memory) or it changes values (it is changed in memory). It is important to appreciate that features such as "Nationality and tempera-

ment of Individual 1" will take different ranges of values for different paragraphs' lexical material. The only significance of the features' values for this analysis is whether or not the values are the same or different in presentation and response. Our analysis here is entirely at the structural level and does not analyse content effects.

We conceive of memory as containing a set of features whose values define each pair of individuals. Our aim is to find a set of features which will adequately model the error data we observe. We assume that sharing features' values makes pairs of individuals more alike; contrasting on features' values makes them less alike. Different features may be more or less important in determining similarity. We make the further standard assumption that errors are more likely to be responses which are *similar* to the stimulus than responses which are *dissimilar*. This approach is no different in principle than the familiar use of confusion matrices to characterise perceptual dimensions of similarity (e.g. Miller & Nicely, 1955). It may be useful to compare this framework to fragmentation. Features are like fragments in that they express partial information. They are unlike fragments in that they may be logically related to each other in complex ways. They are also unlike fragments in that features are always represented by one or other of their values in the memory representation. It is these differences that give rise to redundancy in the representation and formulate retrieval as a constraint satisfaction problem. We return to these points below.

We use regression techniques on the error data to search for and test alternative sets of features for their adequacy in capturing the different frequencies of error types observed. Features are corrupted differently by different error types. The aim is to find a set of features whose pattern of corruption/preservation across the error types can explain the frequencies of those error types. If a feature is corrupted by a given error its inclusion in the set will tend to *decrease* frequencies of that error (it adds a quotient of dissimilarity between stimulus and response). If a feature is preserved by a given error type then its inclusion in the set will tend to *increase* the frequency of that error type (it contributes a quotient of similarity between stimulus and response). The question is whether there is any set of features which will consistently describe the data.

The independent variables of the regression equation correspond to the features. Each of the candidate features' values is computed for the presentation and the recall for each paragraph and they are scored as changed or preserved in memory. They are then collected into the 20 error types shown in Table 2. This is done separately for each match type of stimulus paragraph (there are eight match types of paragraph defined by the two match values—"match" and "mismatch"—on each of the three dimensions other than the introducer which is always mismatched). The

reason for separating the data by match type is that not all errors are possible for all match types of paragraph (e.g. property polarity errors cannot occur on completely mismatched paragraphs). Finally, the proportion of paragraphs in which each feature is correct for each error type is calculated, and these proportions become the values of the independent variables corresponding to the features. For example, for completely correctly recalled paragraphs, all features have a probability of 1.0 of being preserved across presented and recalled models. All errors change some features' values. The details of the classification of error types and the range of candidate features entertained in the search for a regression model of error frequencies are given in Appendix 3.

The dependent variable is a function of error-type frequency. The frequency of occurrence of an error type is adjusted by dividing it by the number of opportunities for its occurrence. This compensates for the unequal sizes of different categories shown in Table 2. The distribution of adjusted frequencies of error types is extremely skewed as a result of the preponderance of responses in which the model is recalled completely correctly. The log of the adjusted frequency, a variable that is nearly normally distributed, is chosen as the dependent variable. The regression coefficients for each feature are then interpretable as weights assigned to features in determining their contribution to the similarity of stimulus to response model.

The model selected from all the candidate features in Appendix 3 is summarised in Table 3. The table gives the regression coefficients and their standard errors for each of the independent variables (features) that were

TABLE 3
Summary of Regression Model Predicting Error Frequencies from Feature Scores

Feature	*Coefficient*	*S.E.*	*P(correct)*
DIMAMAT	1.01	0.15	0.98
DIMBMAT	0.21	0.09	0.93
DIMCMAT	0.49	0.08	0.91
DIMDMAT	0.34	0.08	0.91
NMAT	0.23	0.07	0.83
BD2	0.69	0.11	0.86
AD1	0.37	0.09	0.87
CB1	0.35	0.13	0.90
AC2	0.48	0.08	0.85
B2	0.36	0.13	0.92
BCD1	0.67	0.14	0.85

$R^2 = 0.86$; degrees of freedom = 11,105; intercept = −3.24.

selected by the multiple regression procedure. DIMAMAT to DIMDMAT are features that have a value of 1 if their respective dimensions match and 0 otherwise. NMAT takes a value equal to the number of matched dimensions. The other features take their respective vocabulary items as values, so BD2 takes the vocabulary items of dimensions B and D for the second individual, e.g. "happy, dentist".

The simple model accounts for 86.1% of total variance. All variables contribute significantly to R^2 ($P < 0.01$). The model assumes that features contribute the same coefficient of similarity whenever their values are maintained from presentation to recall, and even with these strong assumptions a reasonably good fit is obtained. Some of the remaining variance is "pure error" (there is variance among the repeated observations of correct recall for example), but in a model with continuous variables it is not easy to estimate the amount of pure error. Nevertheless, the model has significant lack of fit. We will explore some of its successes and failures before discussing its relation to other possible models.

The model successfully accounts for the relations between the frequencies of the main groupings of error types. Figure 1 shows the predicted and observed logged adjusted frequencies averaged over the eight match types. It shows that the relations between the frequency of single, individual polarity, individual polarity + single matching status, complementary double, and homogeneous double errors are well accounted for. There is also considerable success at accounting for the finer detail of the relative frequencies of subcategories within these broad types. For example, the balance of Individual 1 and 2 errors within the several broader categories is quite well predicated.

The model has some minor failings. Completely correct responses are systematically underpredicted, whereas miscellaneous errors are systematically overpredicted. Property polarity errors and their combinations are generally overpredicted whereas individual polarity errors are slightly underpredicted. Multiple polarity errors are generally underpredicted. Analysis of residuals by match type of stimulus model reveals that the completely matched models are fit somewhat worse than the other seven match types, errors on them being generally overpredicted.

The underprediction of completely correct performance, and the overprediction of miscellaneous errors is related to the redundant nature of these representations. Several feature errors are generally involved in making even the least severe response error. Rather large numbers of alternative responses involve similar numbers of feature errors. Few responses actually involve large proportions of wrong features. The underprediction of correct performance and the overpredicition of extreme errors reflect the "error correcting" possibilities of such coding schemes. They may also reflect the fact that there are so far unidentified features

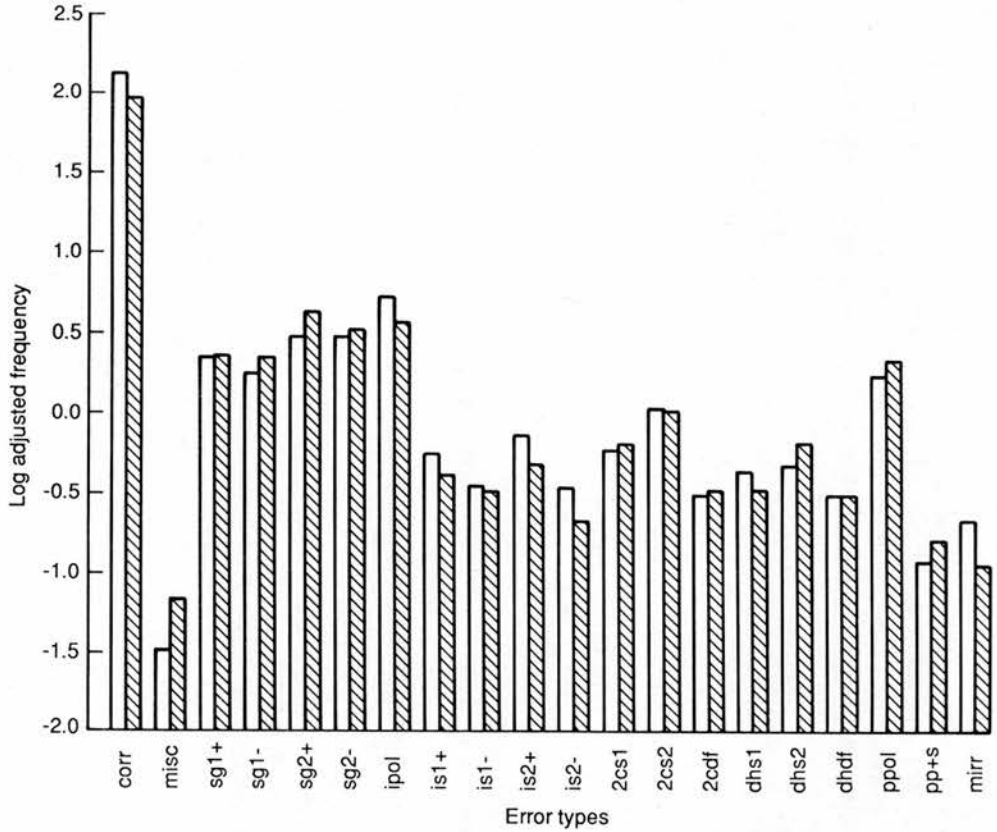


FIG. 1 Predicted and observed log adjusted frequencies. Shaded boxes = Predicted; Open boxes = observed.

which play a minor role in the representation, i.e. any additional features will increase redundancy and accentuate the effects mentioned above. The regression analysis which extracted the model is aimed at finding all the features for which there is significant evidence: There may well be weak features which are present but not sufficiently influential to yield significant improvements of fit. Redundant representations must be expected to have such properties which will inevitably be difficult to investigate.

The overprediction of property polarity errors (especially compared to the opposite tendency for individual polarity errors) probably reflects the special status of matched dimensions. Matched dimensions do not require individuals to be represented *per se*—mere memory of non-occurrence of a lexical items is sufficient: "If nothing red was described, they must be both green." Intuitively, such information is readily available after reading such texts. It would be sufficient to rule out property polarity errors and may well explain their very low observed frequencies. This also helps to explain the observed overprediction of errors for the completely matched models, though there may be additional peculiarities of these models due to their "good figure".

Despite these blemishes, the model shows that the method is capable of extracting a representation composed of diverse logically related features which contribute independently to determining the similarity of stimuli to responses. Some of these features are "fragments" of what was presented, but the majority represent facts that were not presented but which can be inferred from what was presented. The feature set shows no sign of singling out any explicitly presented property as a criterion of identity for the individuals. One might have expected a structure which tied an individual's properties each to an identifier such as the profession. No such pattern emerges, and attempts to fit such models show quite clearly that they are poor fits to the data. The pattern that does emerge achieves complete connectivity—all eight property values are represented in some feature or other. It is noteworthy that there does not seem to be any preponderance of presented intra-individual association (AB, AC, AD) over inferred ones (BC, CD, BD). Nor does the pattern of representing one individual (say I-1) through its intra-individual features and the other (say I-2) through its matching/mismatching with I-1 appear in the analysis, though it is noticeable that I-1 is more integrated than is I-2.

The importance of these regression models at this stage of research lies in their sufficiency. Rather simple models can account for the major patterns in the data. For example, inspection of the data reveals that single mismatching errors occur with individual polarity errors more often than would be expected if the categories were independent. The model predicts the relative frequencies of these two error types, and their co-occurrence accurately. This indicates that the model gives an adequate account of the

interaction between the features that tie properties together within a dimension, and the features that tie properties together within an individual.

We can now see one of our original observations at least partly explained. The main result of log-linear analysis was the observation that intra-individual error correlations were strongest among dimensions B to D and weak between A and the others. This observation is explicable in terms of the current model, and in surprising terms. It does not simply result from different strengths of association between pairs of properties in different structural positions, but is partly mediated through the matching pattern of the three variably matching dimensions (i.e. excluding the introducing dimension).

A subject who forgets the position of an "odd-man-out dimension" (i.e. the mismatching dimension in a description with only one mismatch, apart from the introducer, and the matching dimension in a description where there is only one match) will make a pair of complementary matching errors (see example in Table 1). These pairs of errors can be on distinct individuals, but the model shows how a majority of them fall on the same individual. Here is one source of double errors within an individual involving pairs of non-introducing properties, precisely the type of error revealed as common by log-linear analysis. This explanation does not account for the greatest involvement of dimension D in these errors. Further investigation is required at this point.

The applicability of this framework of memory representations based on independent and redundant features makes possible some rather general observations about the reliability of individual features in memory. Although our subjects' performance is extremely accurate for such a complex timed task, the observation that the underlying representations are so redundant demonstrates that features of the representations must be individually quite unreliable to produce even the low error rates observed. Redundancy allows them to turn in a very respectable performance on a complex task.

READING TIME RESULTS AND DISCUSSION

The reading time data were trimmed to reduce the effect of very long reading times on the means: Values greater than two standard deviations above the mean for each subject for each mode of presentation and sentence position were set to the cut-off value (5.55%, evenly distributed across subjects and sentence positions). The question answering data were trimmed in the same way (5.10%).

The first two analyses were performed to assess the stability of the data over materials and practice. The effects of vocabulary set were assessed by

TABLE 4
Mean Reading Times (sec) as a Function of Individual, Property, Text Mode, and Number of Syllables per Property ("People" Vocabulary Only)

	Individual 1				*Individual 2*			
Property:	*A*	*B*	*C*	*D*	*A*	*B*	*C*	*D*
I × I mode								
2 syllable	1.31	1.65	1.82	3.43	1.64	1.80	2.07	2.40
1 syllable	1.33	1.57	1.79	2.68	1.51	1.73	1.72	2.25
Syllable effect:	−0.02	+0.08	+0.03	+0.75	+0.13	+0.07	+0.35	+0.15
P × P mode								
2 syllable	1.30	1.70	2.19	2.25	1.33	2.07	2.27	2.58
1 syllable	1.35	1.74	1.96	2.32	1.26	1.96	2.19	2.63
Syllable effect	−0.05	−0.04	+0.23	−0.07	+0.07	+0.11	+0.08	−0.05

collapsing across the first and second halves of the experiment. There was no main effect of vocabulary set in any of the text modes (Fs < 1). There were interactions between vocabulary set and sentence—$F(3,48) = 4.15$, $P < 0.015$—in I × I and P × P modes, and between vocabulary set and Individual—$F(1,16) = 4.52$, $P < 0.05$—in MA mode, but none of the differences were significant in *a posteriori* comparisons using Tukey's HSD test. The effects of practice were assessed by collapsing the data across vocabulary sets. Reading times were generally faster in the second half of the experiment—I × I and P × P modes: $F(1,16) = 34.97$, $P < 0.0001$; means first half/second half 2.30/1.64; MA mode: $F(1,16) = 28.01$, $P < 0.0002$; means first half/second half 5.94/4.16—but there were no significant interactions with any other variable. The remaining analyses (except the syllable analysis) collapsed the data across both vocabulary set and practice.

The third analysis was a test of the effect of number of syllables (People vocabulary set only). One-syllable texts were read faster than two-syllable texts—I × I and P × P modes: $F(1,16) = 9.05$, $P < 0.01$; MA mode: $F(1,16) = 20.07$, $P < 0.0005$—but the only interaction involving Syllable was a Mode × Syllable × Individual × Sentence interaction—I × I and P × P modes only, $F(3,48) = 4.41$, $P < 0.01$ (see Table 4).

At most sentence positions there is no evidence of articulatory rehearsal slowing reading time. Baddeley et al. (1975) observed a mean of about 0.2 sec per syllable rehearsed. With this figure as a rough estimate, our data suggest that an articulatory rehearsal review of the first individual in I × I texts may take place at sentence four (i.e. Individual 1, Dimension D), and in MA texts both individuals may be rehearsed in this way. The extra four

TABLE 5
Mean Reading Times (sec) as a Function of Individual, Property, and Text Mode (Both "Object" and "People" Vocabularies)

	Individual 1				*Individual 2*			
Property:	*A*	*B*	*C*	*D*	*A*	*B*	*C*	*D*
I × I mode	1.36	1.57	1.82	3.16	1.63	1.74	1.94	2.45
P × P mode	1.39	1.62	2.04	2.56	1.36	1.95	2.36	2.73

syllables involved in the specification of an individual in two-syllable word texts add 0.75 sec to reading time at sentence four of I × I texts, 0.98 sec and 0.99 sec to sentences one and two respectively of MA texts. These figures are wholly consistent with the existence of an articulatory rehearsal process at these text positions which operates as some sort of auxiliary memory. This constitutes an interesting observation of this phenomenon during a task dominated by semantic processing. Equally, it is clear that the semantic ordinal effect is not due to increasing articulatory rehearsal with an increase in the number of attributes known. Nor is articulatory rehearsal the basis for memory at other positions in these texts.

The main analysis was a three-way analysis of variance for Mode (I × I versus P × P) × Individual × Property. Table 5 displays mean reading times by mode, individual, and property. The only significant main effect was Property—$F(3,48) = 26.60$, $P < 0.0001$—and the only significant interaction was Mode × Individual × Property—$F(3,48) = 5,60$, $P < 0.0025$.

In MA mode, the second sentence (4.80 sec) was read faster than the first sentence (6.22 sec). This difference was not significant—$F(1,16) = 3.18$, $P < 0.095$). These times compare with accumulated reading times (in I × I mode) of 7.91 sec for the first individual and 7.76 sec for the second individual.

Averaging across individuals, the mean reading time for MA sentences (5.5 sec) falls midway between the time for the final sentences about individuals (2.8 sec), and the accumulated time for four properties of individuals (7.8 sec) in I × I texts. If the MA time were equal to the lower of these two figures, that would indicate that all representational construction awaited the arrival of the fourth property, and that previous activity was maintenance rehearsal of some sort. If the MA time were equal to the greater of these two figures, that would indicate that all the representational construction carried out during sentences 1 to 3 of I × I texts was carried forward to sentence four. What extra overheads there are in I × I texts (16 words instead of 7 to read, three more bar presses, etc.) would

tend to bring the relevant estimate of the accumulated time nearer to the MA time observed. We can therefore conclude that the balance of time consumed by maintenance as opposed to constructional activities falls more toward construction. This method of estimation is crude but it agrees with the localised articulatory rehearsal effects observed.

The above analyses replicate Stenning (1986) in establishing the semantic ordinal effect as a phenomenon which remains stable over different experiments, vocabulary sets, levels of practice, and modes of text presentation. The semantic ordinal effect is seen most clearly in the I × I mode of presentation: Reading times increase through the first four sentences as properties are added to the first individual, decrease almost to baseline at the introduction of the second individual, and then increase as properties are added to the second individual. When the sentences of a P × P text are sorted into property (rather than temporal) order, as in Table 3, a very similar pattern of reading times is observed. Processing proceeds oriented towards what is known about the individual currently referenced.

In what follows, we explore the idea that the semantic ordinal effect reflects chiefly the construction of representations for referenced individuals. Reading time is not merely a function of the number of properties known of the referenced individual. Inspection of the data shows that mismatched properties generally took longer to read than matched properties. The recall error modelling has already shown that matching information is an important part of subjects' representations. In order to explore the degree of modularity of the processes which take up reading time, and to see how they could be related to the representations revealed by error analysis, we developed a multiple regression model which used the match/mismatch structures of the texts to predict reading times. The development of this model is summarised below. Our aim is to factor reading time into several distinguishable parts each of which would occur whenever its conditions were fulfilled, and which would be the same whenever it occurred in a text. If this could be achieved, these parts could then be interpreted as times occupied by invariant processes.

The simplest model of all would simply count the number of properties known about the referenced individual and add a constant increment of time for each. Such a model is too simple for two reasons. First, the matching status of properties makes a difference and, secondly, the time taken is not linear with the number of properties. So the model distinguishes the matching status of properties, and has a separate "dummy" variable for each number of a given status of property in order to allow the data to indicate the shape of the functions of load imposed by different numbers of various sorts of items. The shape of these functions will then be used to constrain theories of what processes underlie these reading times.

In the present experimental texts, assignment of a property on a given dimension (e.g. colour) to an individual always occurred first for the first individual, and subsequently for the second individual (but see Stenning et al., 1987, for an experiment which manipulates sequence of references more generally). With respect to the corresponding property on the other individual, a property can match (e.g. when both objects are red), mismatch (e.g. one object is red and the other green), or be indeterminate (e.g. when the colour of the other object has not yet been specified). Three load factors (called MATLOAD, MISLOAD, and NEUTLOAD, respectively) represented the accumulated number of each of these three sorts of properties at a given point in the text. Dummy variables were defined which uniquely identified each level of each of these factors.

In addition to these processing loads, there might be processes which occurred locally with certain textual events such as changing reference or detecting/encoding the existence of a mismatch. A distinction is made between the processing time occurring on first learning that a property is mismatched, and the processing time which results from the knowledge of a mismatched property of the referenced individual. The latter will recur when the individual is referred to again; the former will not.

Separate regression models were developed for I × I and P × P texts, on the grounds that there might be strategic differences between the two types of text, but these models were so similar that we present a general model for data from both text types. Definitions of the variables are as follows:

1. NEUTLOAD is the number of properties on the referenced individual which cannot be assigned as matches or mismatches with the background individual. This factor was expressed as dummy variables NEUT1, NEUT2, NEUT3, NEUT4. Each had value 1 if NEUTLOAD's value corresponded to its number, otherwise its value was 0.
2. MATLOAD is the number of matches on the referenced individual. It does not come into effect until a mismatch (other than the mismatch of the introducers) has occurred; it then takes full retrospective effect. We explored the alternative straightforward definition where MATLOAD comes into effect immediately. The current definition fits the data better, and has some intuitive justification since it is only when a second mismatch arises that readers need to process both individuals as such. This factor was expressed as dummy variables MAT1, MAT2, MAT3 in the same manner as NEUTLOAD.
3. MISLOAD is the number of mismatches on the referenced individual. This factor was expressed as dummy variables MIS1, MIS2, MIS3, MIS4 in the same manner as NEUTLOAD.
4. LOCALMIS is a binary variable which is 1 when a mismatch is

detected, and 0 otherwise. It does not count the detection of the predictable mismatch of the introducing properties.

These definitions are given in full for all text modes and match types in Appendix 4. The resulting regression model is presented in Table 6.

All variables contribute significantly to R^2 ($P < 0.005$). Pure error accounts for 87.66% of the total variance. The regression model accounts for 97.07% of the remaining variance, leaving 0.36% lack of fit. Figure 2 shows the observed and predicted reading times at each sentence position for each match type. The +/− notation should be interpreted as defining the matching/mismatching of the three non-introducing dimensions, e.g. ++− refers to the structure where the mismatched introducing dimension is followed by two matching dimensions and one mismatching one. The impact of match type is clearly visible in the processing of the second individual.

Each of the values for the three sorts of load occurs at several different points in the various texts, and with different combinations of values of the other two load variables. The model's good fit indicates that the loads do indeed accumulate without interacting with each other and not as a function of their history in the text.

What interpretation should be placed on the load variables of this model? Two questions arise: First, how are we to interpret the relations between the three load variables, i.e. the fact that for all but one load level, MISLOAD contributes a greater time than does NEUTLOAD, and NEUTLOAD, in turn, greater than MATLOAD? Secondly, how are we to interpret the shapes of the functions?

TABLE 6
Summary of Regression Model Predicting Reading Times from Match Structure

Variable	*Coeff(sec)*	*Standard Error*
Intercept	1.09	0.066
NEUT1	0.28	0.049
NEUT2	0.47	0.092
NEUT3	0.73	0.092
NEUT4	2.07	0.092
MAT1	0.42	0.063
MAT2	0.58	0.090
MIS1	0.46	0.062
MIS2	0.79	0.076
MIS3	1.15	0.090
MIS4	1.37	0.150
LOCALMIS	0.19	0.065

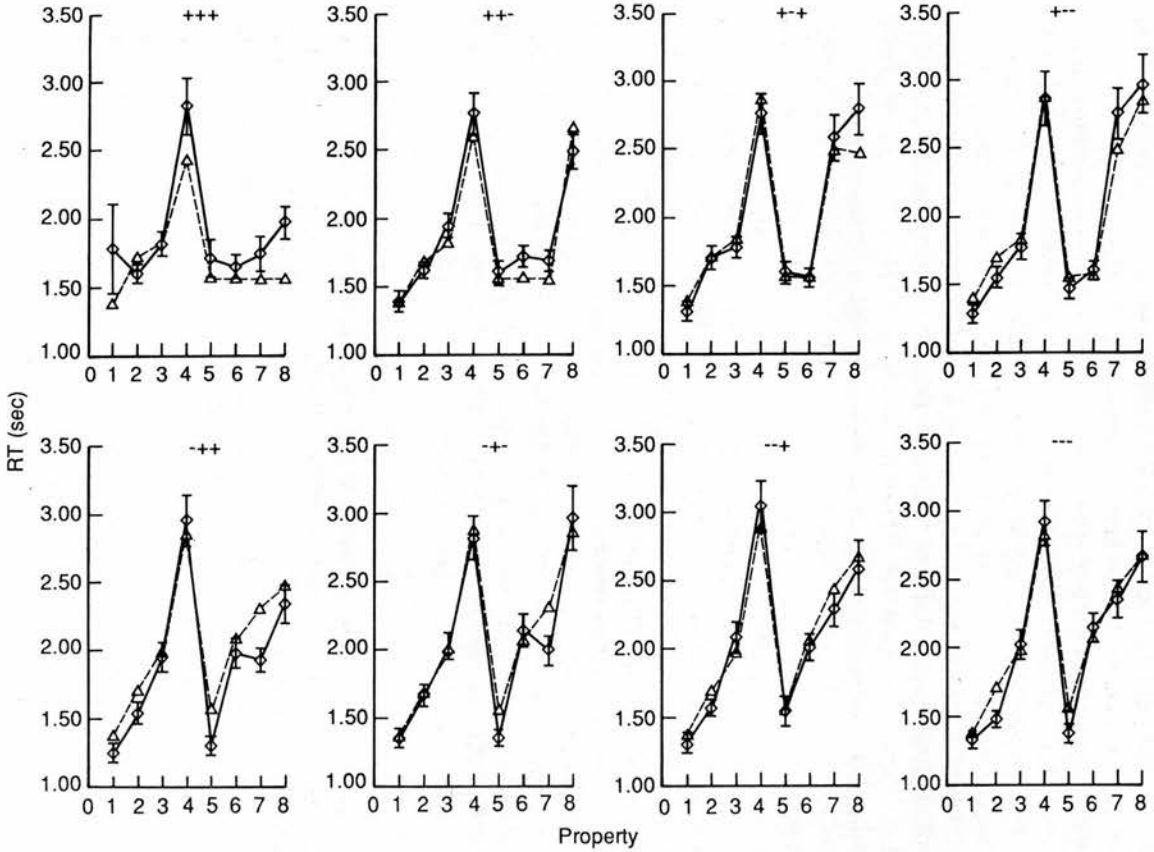


FIG. 2 Predicted and observed reading times for each matchtype. ◇—◇, Observed; △— —△, predicted.

On the first question, we interpret the relations between the three functions not as reflecting different sorts of processes but in terms of the number of occasions on which the same sorts of processes need to be brought into play. It is only when individuals contrast on a property other than the introducer that the need arises to treat them as individuals, in the sense that we need to represent relations between the several properties of an individual. If there is a bishop and a chef, and they are both thin and both Polish and both tall, we simply need to remember the three properties ("thin", "Polish", "tall" as opposed to "fat", "German", "short"). If, however, one is German and one is Polish, then we must remember which is which—we must associate, say "Polish" and "bishop", "German" and "chef". Another mismatch between "tall" and "short" compounds the problem, and so on. The number of alternatives rises exponentially.

In all texts, as MISLOAD increases, the number of intra-individual associations required increases. What about NEUTLOAD? The values two to four of this variable are estimated exclusively from I × I mode data. In reading the first individual in an I × I text it is unpredictable whether intra-individual associations will be necessary, because the match structure is unknown, and we assume that readers form all or most of them in case they are necessary. The fact that NEUTLOAD contributes less than MATLOAD at the first three points may be because the reader gambles on not needing all such links, or because in these predictable texts, the subject delays some of the processing of these links until sentence four. The total time under the two functions is 3.55 sec for NEUTLOAD and 3.77 sec for MATLOAD, consistent with the latter explanation. The fourth sentence of I × I texts also has its reading time lengthened by an articulatory rehearsal process (see above).

There is only any solid evidence for MATLOAD contributing to reading time after the occurrence of an unpredictable mismatch, and this again suggests it is the formation of intra-individual associations which is accounting for processing load. Even after such a mismatch, encoding associations between properties on matched dimensions and other properties is not strictly necessary, but adds redundancy.

Altogether, the data suggest that it is when intra-individual associations are formed that processing load increases. We propose that match structure affects processing load not through the necessity of itself being represented, but through its effects on which intra-individual associations are useful. It is the intra-individual associations that are novel with each new paragraph's lexical material.

The second question, what gives rise to the slopes of the three load functions, reveals a tension between the cumulativeness of the times and the idea that representations once constructed are carried forward and do not need to be constructed again. The evidence from the MA texts and the

localised articulatory rehearsal effects suggest that mere maintenance consumes little time. The cumulative functions suggest that either the number of *new* intra-individual associations to be formed increases linearly with number of properties known of an individual, or the difficulty of forming associations constrained by more properties is greater and the time necessary is longer.

If we take the regression model of recall as a guide to which intra-individual associations are formed, we can count the number of new associations completed by each sentence. The resulting functions are much flatter than the observed load functions, particularly early on in the specification of the individual. On balance, the evidence suggests that the additional constraints imposed by more properties must slow down the search for acceptable associations. We return to this matter below.

QUESTION ANSWERING TIME RESULTS

Analysis of question answering times by mode of presentation, vocabulary set, practice, individual, yes/no, and correct/incorrect failed to show any consistent differences between these conditions.

Error rate was 11% overall, with tendencies (reliable across mode, vocabulary set, and first/second halves of the experiment) for more errors on the second individual (14%) than on the first individual (8%), and more errors for "no" responses (14%) than for "yes" responses (9%) (Mann-Whitney text, $P < 0.05$ in both cases).

This is the most widely used of our three subtasks in probing the structure of memory representations. The fact that it yields no useful data in this context is probably a result of several circumstances. First, the effects are small compared to those in the other two subtasks. Secondly, in retrieval tasks, subjects are tested in circumstances where they are trained on the information up to criterion, and then readied for a purely testing phase, whereas our subjects have to switch tasks each paragraph. Finally, our readers are quite successful at titrating their reading times to secure error rates even across the properties in a text. In fact, their learning is probably more even across parts of the structures than when they are trained to a criterion, but their degree of learning is undoubtedly lower than in the standard tasks.

GENERAL DISCUSSION

Our task has similarities both to those used by researchers who study memory for interpreted material and to rote verbal learning tasks. At first it may appear more similar to the latter, i.e. the subject is faced with a sequence of items, each containing one novel content word. Hardly an

inference is required, only verbatim recall. There is no plot. No implicit background has to be conjured to bridge gaps. But these appearances are misleading. Subjects learn the material by bringing to bear their knowledge of people (or shapes), and they do so rapidly and accurately. Although this material might be expected to be maximally mutually interfering, subjects suffer little proactive interference. Just introducing the possibility of interpreting items as properties of individuals radically changes the nature of the task. It is like one of the subtasks faced by people reading more natural texts: They must derive a mapping of properties onto individuals and remember this mapping. This task of representing an extensional framework for a text is our concern.

The reading times show how tightly processing is organised around individuals. Whether the non-referenced individual has already been completely described, or is only as yet partially specified, the amount known about that individual has no impact on time to process the current new information about the currently referenced individual. On the other hand, the amount already known about the currently referenced individual strongly determines processing time, and as that amount of knowledge increases, so does processing time. Some of this extra processing time is related to maintenance activities, but a good proportion is taken up with the construction of more durable representations.

Little of the increase of time with amount known can be accounted for by syllabic rehearsal, though that does not mean that rehearsal does not take place. It is often possible to *hear* subjects rehearsing. The evidence from the processing of P × P texts suggests that if rehearsal takes place, items are being re-entered into the articulatory loop (Baddeley, 1986) from some representation other than the articulatory/acoustic store: Too much time and material intervenes between successive rehearsals of the same individual, especially late in the texts. Stenning et al. (1987) further analyse syllabic rehearsal in this task and conclude that rehearsal may be being used to facilitate the imposition of the semantically based grouping which is demanded by the task.

The increase in reading time with increasing knowledge about the currently referenced individual is not to be accounted for by increased rehearsal but by increased semantic processing. What construction consumes this extra time? The analysis of errors indicates that the representations underlying memory in this task are divided into sets of features which contribute independently to the retrieval of a response, but which are logically interrelated in complex ways. The net effect is a highly redundant representation. How are the attributes of individuals bound together in these representations? Not by each feature containing an explicitly mentioned identifying property. In the regression model offered here, binding occurs through implicit contextually identifying properties (order of men-

tion). Some of the features (the matching features and NMAT) do not even contain any reference to individuals but play their part in achieving binding of properties to individuals none the less. It is a question currently being actively pursued whether even the residual contextual referring elements in these representations are necessary. Stenning et al. (1987) show even more radical models, in which purely existential statements about the instantiation of conjunctions of properties can explain the observed data from a somewhat generalised version of this task. At any rate, the general answer about how binding is achieved seems to be that it is achieved through the synthesis of information contained in many independent and redundant representations.

Jones' work first produced evidence that representations underlying memory for individuals were composed of sets of independent fragments. His material is designed so that each fragment of two or more properties is a unique cue to its individual. The situation Jones studied differs from the present one in that rather few fragments need be represented in memory to account for the performance levels observed, and inferences are not required to synthesise fragments into whole descriptions. The binding problem is solved in the fragmentation theory by structure internal to fragments.

In the case of the representations revealed in the model of errors developed here, the synthesis of feature values into an integrated description depends on complex inferences—the cost of redundancy combined with fragmentary representations is complexity of retrieval inference. This is particularly so when corruption makes the state of the representation inconsistent with any interpretation, and a best-fit must be found to multiple constraints. This need for an inference mechanism thrown up by the discovered representations has lead us to explore PDP systems as devices for making these retrieval inferences. Stenning and Levy (1988) show that a PDP system can be fabricated directly from the regression model of errors like that developed here, and that this PDP system produces human-like errors when synthesising descriptions from noisy sets of features.

We see our theory as a first step toward explaining why human solutions to the binding problem are content-dependent. Although our analysis here is in terms of features corresponding to propositions in which properties are structurally conjoined, we do not think of their implementation as being in terms of a compositional representation like a logic. A feature such as "bishop and Polish" might well be implemented in memory by a recruited association such as "catholic". Given a menu on which "Polish", "Swiss", "bishop", and "dentist" appeared, the presence of an association like "catholic" would be sufficient to implement the fact that there was a Polish bishop. Of course, such recruited associations might be more in-

effable than such neatly lexicalised properties. It is the recruitment of these sorts of associations which we assume is the dominant process which accounts for the reading times observed. We are currently pursuing these issues of content by analysing properties of subjects' beliefs about the pairs of people they read about, and the impact of those beliefs on their memories.

Revised manuscript received 29 April 1988

REFERENCES

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*, 249–277.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, Mass.: Harvard University Press.

Anderson, J. R. & Bower, G. H. (1973). *Human associative memory*. Washington D.C.: Hemisphere.

Baddeley, A. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575–589.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1974). *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass.: MIT Press.

Hirst, G. (1981). *Anaphora in natural language understanding*. Berlin: Springer-Verlag.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge: Cambridge University Press.

Jones, G. V. (1976). A fragmentation hypothesis of memory: Cued recall of pictures and of sequential position. *Journal of Experimental Psychology, 105*, 277–293.

Kieras, D. E. (1981). Component processes in the comprehension of simple prose. *Journal of Verbal Learning and Verbal Behavior, 20*, 1–23.

McClelland, J. L. & Rumelhart, D. E. (Eds) (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, Mass.: MIT Press.

Miller, G. A. (1956). The magical number seven plus or minus two, or, some limits on our capacity for processing information. *Psychological Review, 63*, 81–96.

Miller, G. A. & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27*, 338–352.

Newell, A. (1973). Why you can't play twenty questions with nature and win. In W. G. Chase (Ed.), *Visual information processing*, p.10. London and San Diego: Academic Press.

Rumelhart, D. E., Lindsay, P. H., & Norman, D. A. (1972). A process model for long-term memory. In E. Tulving & W. Donaldson (Eds), *The organisation of memory*, pp. 197–245. London and San Diego: Academic Press.

Sanford, A. J. & Garrod, S. C. (1981). *Understanding written language*. Chichester: John Wiley.

Schank, R. C. & Abelson, R. (1977). *Scripts, plans, goals and understanding*. Hillsdale, N.J.: Lawrence Erlbaum Associates Inc.

Stenning, K. (1978). Anaphora as an approach to pragmatics. In M. Halle, J. Bresnan, & G. A. Miller (Eds), *Linguistic theory and psychological reality*, pp. 162–199. Cambridge, Mass.: MIT Press.

Stenning, K. (1986). On making models: A study of constructive memory. In T. Myers, K.

Brown, & B. McGonigle (Eds), *Reasoning and discourse processes*, pp. 165–185. London and San Diego: Academic Press.

Stenning, K. & Levy, J. (1988). Knowledge-rich solutions to the 'binding problem': Some human computational mechanisms. *Knowledge Based Systems, 1*, 143–152.

Stenning, K., Patel, M. J., & Levy, J. (1987). *The 'Binding Problem' in human memory: Some effects of referential discontinuity on the construction of representations for individuals.* Technical Report, Centre for Cognitive Science, Edinburgh University.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Oxford University Press.

APPENDIX 1: VOCABULARY

Dimension A	*Dimension B*	*Dimension C*	*Dimension D*
People vocabulary set			
nurse/priest	French/Greek	young/old	tall/short
judge/monk	Welsh/Swiss	fat/thin	strong/weak
vet/chef	Dutch/Czech	rich/poor	sane/mad
doctor/vicar	German/Spanish	clever/stupid	friendly/hostile
teacher/bishop	Chinese/Polish	hungry/thirsty	happy/gloomy
dentist/baker	Swedish/Russian	greedy/clumsy	daring/timid
Objects vocabulary set			
circle/square	black/white	old/new	thick/thin
triangle/oval	red/green	hard/soft	deep/shallow
rectangle/ellipse	yellow/blue	rough/smooth	hollow/solid
beam/block	silver/gold	wet/dry	large/small
cylinder/pyramid	bright/dim	hot/cold	long/short
disc/cube	shiny/dull	light/heavy	wide/narrow

APPENDIX 2: DESCRIPTIVE ANALYSES OF RECALL ERRORS

Table 7 shows percentages of error by position of presentation and position of recall. Cases in which a single property error is made on an individual are distinguished from multiple property errors.

Errors were more correlated with recall position than with presentation position. This could be because the subject chooses to recall what is best known first, or it could be because recalling the first individual interferes with memory of the second. Results from other experiments suggest the latter interpretation (Stenning et al., 1987). This effect of recall order was strongest for multiple errors. There was a tendency, mentioned above, to recall the individuals in their order of presentation (1071 in order, 466 in reverse order).

The following analyses are organised into those that centre on properties—that is on correlations between performance on the same property of the two individuals—and those that centre on individuals—that is on correlations between performance on the different properties of a single individual.

TABLE 7
Percentage of Single and Multiple Errors as a Function of Stimulus Position and Order of Recall ($N = 1537$)

	Single Error		*Multiple Error*	
Stimulus Position	*First*	*Second*	*First*	*Second*
First recalled individual	16.0	14.0	2.1	3.6
Second recalled individual	16.0	18.0	8.4	8.1

Property-oriented analysis

On a given property dimension, a subject could either make no errors, an error on Individual 1 but not on Individual 2, an error on Individual 2 but not on Individual 1, or an error on both. In addition, a property dimension may be matched or mismatched in presentation. Table 8 shows the percentages of error classified in this way. The introducing property is always mismatched and so is treated separately: The other three dimensions are collapsed because their patterns are similar.

For all properties there was a tendency for second individual errors to exceed first individual errors (first and second are defined by recall order). There was also a general tendency for the errors to be correlated: Double errors were much more common than would be expected, especially on mismatched dimensions.

The nominal property behaved differently from the other properties, even the other mismatched properties: The nominals show an especially high correlation of errors on one individual with errors on the other individual. This may be because subjects realised that they all mismatched.

Properties B to D showed evidence of strong correlation of errors between Individual 1 and Individual 2, though less strongly so than the nominal properties. The effect of mismatching properties was to accentuate this correlation.

Individual-oriented analysis

The percentages of errors for each dimension and each individual are shown in Table 9. They are broken down into responses in which only a single property was in error, and responses in which multiple properties were wrong.

TABLE 8
Percentage of First, Second, and Both Object Errors as a Function of Matched and Mismatched Properties

	Property A[a]		*Properties B–D*	
	Matched	*Mismatched*	*Matched*	*Mismatched*
1st object	—	0.72	1.55	2.45
2nd object	—	1.56	6.79	6.40
Both objects	—	4.75	1.85	4.42
Correlation coeff.	—	0.80	0.30	0.47

[a]Introducing dimension always mismatches.

TABLE 9
Percentages of Single and Multiple Errors Across Properties within Individuals

	Property			
	A	*B*	*C*	*D*
Single errors				
Individual 1	4.68	2.99	3.51	4.03
Individual 2	3.71	4.03	5.27	4.36
Multiple errors				
Individual 1	0.78	1.24	1.63	2.02
Individual 2	2.67	4.55	5.14	5.66

Single errors were the most frequent categories of errors on both individuals. They did not show any consistent or pronounced trend across the four properties. This suggests that subjects were sucessfully titrating their reading time across the sentences of the paragraph, and holding these errors constant.

However, Table 9 shows that multiple within-individual errors are not randomly distributed across the properties: There is a tendency for property D to be included in multiple errors. It is also noticeable that the increase in errors from the first to the second recalled individual includes a disproportionate increase in multiple errors.

In the log-linear model analyses that follow (see Bishop, Fienberg, & Holland, 1974), data from Individual 1 is treated separately from that from Individual 2. For Individual 1, the most parsimonious adequate model is (DB, DC, A; $P = 0.32$). For Individual 2, the most parsimonious adequate model is (DC, DB, CB, BA; $P = 0.11$).

In the organisation of Individual 1, there were strong correlations between the fate in recall of property D and properties B and C. In the organisation of Individual 2, this pattern of correlations between D and the other properties persisted, but additional correlations appear, namely CB and BA. All terms in both these models represent positive correlations: Correct performance on one property is positively correlated with correct performance on the other. The differences between Individual 1 and Individual 2 are due to the increase in multiple error rates for the second recalled individual that were noted in Table 7. Note that the best model for Individual 1 contains a proper subset of the terms in the best model for Individual 2. The difference between multiple error rates for the two individuals appears to be an increase in the number of the same sort of multiple errors.

APPENDIX 3: ERROR-TYPE CLASSIFICATION AND FEATURES OF THE RECALL ERROR REGRESSION MODEL

A classification of error types

The simplest type of error is a single property wrong. Single errors can occur on Individual 1 or Individual 2, and on matched or mismatched dimensions. Note that single errors always alter the matching status of the affected property dimension—either matched to mismatched, or mismatched to matched.

If two properties are incorrect in a response, there are several ways in which the two properties may be related. If two errors occur on a single property dimension they will be termed *polarity errors* (because the polarity of either the individuals or the properties has been switched). Polarity errors will bç dealt with below. If two errors are on separate dimensions, then they may be either *homogeneous* (both change matched dimensions to mismatched ones, or both change mismatched dimensions to matched ones), or *complementary* (one of each type of change). At the same time, for both homogeneous and complementary error pairs, both members of the pair may occur on Individual 1, both on Individual 2, or one on each individual.

Polarity errors, in which two errors occur on a single property dimension, never disturb the matching status of their dimension. They are of two types. If the dimension is matched (say both objects red), then to recall them as both wrong, and hence still matched (both green), will be termed a *property polarity* error. If the dimension is mismatched (a red and a green), then to recall them as both wrong, and hence still mismatched, will be termed an *individual polarity* error. As with error of matching status, polarity errors can occur in multiples (double individual polarity, double property polarity, or mixed polarity). Polarity errors can also occur in combination with single errors on other dimensions.

Finally, most possible error types fall into none of these categories, but also hold little obvious theoretical interest. They are rare in the data, and will be resigned to a miscellaneous category.

Candidate features

The following features were entertained in the search for an adequate regression model of error frequencies. Features integrating the properties of a single individual—all possible combinations (from features containing a single dimension to features containing quadruples of dimensions) were constructed. Features integrating the property dimensions—a feature was defined for each of the four dimensions, each feature taking the values "match" or "mismatch". Meta-matching features—a feature called NMAT was defined whose value was the number of matched dimensions. Any two models which have the same number of matched dimensions share values on this feature, whereas any two that have a different number of matched dimensions contrast with regard to this feature.

APPENDIX 4: DEFINITIONS OF VARIABLES FOR THE REGRESSION MODEL

	Sentence[a]							
	1 N+−L	*2* N+−L	*3* N+−L	*4* N+−L	*5* N+−L	*6* N+−L	*7* N+−L	*8* N+−L
I × *I texts*								
+++	1000	2000	3000	4000	0010	0010	0010	0010
++−	1000	2000	3000	4000	0010	0010	0010	0221
+−+	1000	2000	3000	4000	0010	0010	0121	0220
+−−	1000	2000	3000	4000	0010	0010	0121	0131
−++	1000	2000	3000	4000	0010	0021	0120	0220
−+−	1000	2000	3000	4000	0010	0021	0120	0131
−−+	1000	2000	3000	4000	0010	0021	0031	0130
−−−	1000	2000	3000	4000	0010	0021	0031	0041
P × *P texts*								
+++	1000	0010	1010	0010	1010	0010	1010	0010
++−	1000	0010	1010	0010	1010	0010	1010	0221
+−+	1000	0010	1010	0010	1010	0121	1120	0220
+−−	1000	0010	1010	0010	1010	0121	1120	0131
−++	1000	0010	1010	0021	1020	0120	1120	0220
−+−	1000	0010	1010	0021	1020	0120	1120	0131
−−+	1000	0010	1010	0021	1020	0031	1030	0130
−−−	1000	0010	1010	0021	1020	0031	1030	0041

[a]N, NEUTLAND; +, MATLOAD; −, MISLOAD; L, LOCALMIS.

Knowledge-rich solutions to the binding problem: a simulation of some human computational mechanisms

Keith Stenning* and Joe Levy†

The binding problem, how properties are represented as belonging to individuals, is identified as a severe problem for human memory, for which the memory adopts knowledge-rich solutions. It is argued that it is the nature of these solutions that endows human memory with many of its positive properties, particularly rapid retrieval on the basis of unreliable search clues. Parallel Distributed Processing (PDP) systems offer some insight into how human memory systems may work, as they also have to solve the binding problem by knowledge-rich methods. Experimental analysis and statistical models of Memory for Individuals Task (MIT) are presented, which provide evidence that the memory representations underlying human performance consist of sets of existential facts containing no referential terms. It is shown that the proposed representations can be incorporated directly into a PDP simulation of the inference from representation to response, and that the resulting system produces human-like errors when subjected to noisy input. The PDP simulation captures some of the asymmetries between stimulus and response which the statistical model cannot.

Keywords: binding, memory, PDP system, knowledge-rich, human memory, representations

The binding problem is something that knowledge representation systems have to solve: how are attributes represented as being possessed by one individual rather than another? It is a problem which is so easily solved in a variety of elementary ways by conventional von Neumann computers that it hardly seems the stuff with which knowledge-based systems are especially concerned. This work begins from a psychological perspective, asking how human memory solves the binding problem. It is of relevance to those who design knowledge-based systems for two reasons. First, it is important for the designer to understand human memory limitations so that machines can avoid playing on human weaknesses: systems must respect the capacity of human working memory. Second, and more ambitiously, it is important to understand the nature of the human solutions to this problem because they expose what is definitive about human memory — that its processes and structures are knowledge-rich. This knowledge richness gives rise to its elementary limitations, but also to its phenomenal abilities. It is this ability to rapidly retrieve appropriate information from an enormous database, and on the basis of partial and unreliable information, which feeds all other cognitive abilities: speech perception, language, vision, etc. The argument presented here is that to simulate these memory abilities the engineer may have to forego elementary structural solutions to the binding problem, and adopt something more like the solution we see in human memory.

To suggest intellectual traffic in this direction is to reverse the most common pattern of trade. Many problems in cognitive psychology owe the impetus of their investigation, if not their discovery, to the engineer's problems. For example, the *anaphor resolution* problem (how to choose between several possible antecedents for a pronoun) is a problem closely related to the binding problem which has received much attention from both computer scientists and psychologists (see Reference 1 for a review). The anaphor resolution problem is severe for the machine because human beings solve it (largely without noticing) by bringing to bear rich knowledge of their language and the world. But the binding problem has the opposite characteristics: it is easy for the machine to solve by structural methods without recourse to knowledge of content, but for the human being binding can pose severe memory problems. This reversal of the direction of traffic means that the psychologist may have to persuade the engineer to take on a problem which

Centre for Cognitive Science, †Department of Psychology, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK

0950–7051/88/030143–10 $03.00

it was supposed had been solved. However, perhaps the engineer will not take too much persuading: it is commonplace in artificial intelligence that many problems have been taken as far as they can using syntactic techniques, and that ways of mobilizing rich knowledge are the only long term avenue for progress.

The investigations of human memory reported here are experimental psychology statistical modelling and PDP simulation, rather than the design of software. Experimental analysis in this degree of detail requires tightly constrained and often-repeated situations which may seem divorced from both the applications that concern the human computer interface expert, and from the knowledge storage and retrieval systems which we might want to implement in a machine. Their justification comes from the knowledge representation issue they address.

WHEN IS BINDING A PROBLEM FOR PEOPLES' MEMORY?

Psychologists since Tulving[2] have conventionally distinguished between episodic and semantic memory, a most unfortunate terminology; episodic memories are generally thoroughly semantically interpreted. The distinction is between memory for specific personal experiences as opposed to memory for generalized knowledge. We remember, as part of our generalized memory, that aardvarks are animals and Pythagoras was a Greek mathematician, and typically we have no memory of the *experience of learning* these facts, even in the cases in which there was a single learning experience. If we do remember such an experience, then that is an episodic memory. One does not have to believe there are two 'memories' in two parts of the brain to believe that these types of information require distinction. They differ sufficiently in the informational characteristics to require distinct treatments in theories of memory. The human ability to remember and 'revisit' personal experiences is phylogenetically recent, and must rank as one of the species' outstanding cognitive attributes. What is crucial to this claim is that human beings interpret present reminiscences as having reference to earlier times. The dog that dreams may or may not re-experience some previous experience, but it does not interpret the dream as having reference to another time.

In an episodic memory we fix a unique experience in terms of its combination of properties, *at one go*. When we observe that our ability to do this is dependent on the content of the experience, we are saying that episodic memory is mounted on top of our general memories. It is this relation between episodic memory and general knowledge which particularly concerns us here. The binding problem addresses the representation of combinations of properties, and how the combinations of properties experienced are distinguished from other combinations of the same properties. Binding presents a problem for human memory when many combinations of properties are plausible, and that happens when general knowledge is not sufficient to rule out many combinations. So binding is an active problem for memory when an experience has one out of many plausible combinations of properties.

It is an obvious property of human memory (one that has been evident to psychologists for as long as memory has been studied) that the general knowledge subject bring to the laboratory is an important determinant o how memorable new information presented to them wil prove to be. Bartlett[3] made this point most forcefully and since then those who have been concerned with the role of meaning in memory have built theorie around the fact that general knowledge reduces th amount of information which we have to remembe because it introduces redundancy. Told that Bill wen into a restaurant, bought a hamburger, left a tip fo the waitress and left, we do not have to discriminat this combination of events in memory from one in whic a hamburger bought a waitress from a tip and left restaurant in Bill. We do not have to remember th combination because we know it already as a specia case of a general 'script' (e.g. see Schank and Abelson[4])

However, general knowledge does more than reduc the information load, and there is more to memory thar what we already know. Unfortunately, those psycho logists who have most stressed the learning of nove combinations of items in episodic memory, followin Ebbinghaus' lead[5], have felt it necessary to attempt t minimize the effects of meaning in their materials b using the learning of unstructured lists of nonsens syllables, numbers or words as their chosen task. Thes tasks are what give rise to the particularly stark contras between human and machine memory. Given lists o random digits, a person can repeat back perhaps sever items, and then only if allowed to do so immediately.

The program of eradicating meaning from experi mental materials has had unfortunate effects, becaus meaning plays rolls in memory other than the introduc tion of redundancy. Suppose we learn that Bill was wear ing a blue shirt when he went into an Italian restaurant and ordered a vegetarian lasagna from a waiter in a red shirt, whereas Fred wore a yellow shirt and ordered canelloni. If we are then asked about the combinations of meals and shirts we have a binding problem. Nothing in general knowledge prohibits other combinations of shirts and meals: we do not have a script for blue shirted vegetarian lasagna eating. Nevertheless, these combinations are much easier to remember than equivalently informationally rich combinations of digits. Although these combinations are not either more or less likely than each other, they can be imbued with meaning which 'fixes' them in memory. What we require is a theory that will explain this role of meaning in determining memory for combinations of values of orthogonal variables.

The other side of the coin of human memory from the pathetic seven arbitrary items capacity of immediate memory, is retrieval ability. The capacity of human long term memory is inestimable, but large by the standards of even large computer databases. Despite the amount of memory to be searched, we can assess much of what is accessible from a given context of retrieval within seconds rather than minutes, and can do so even when the cues we are given are unreliable (see Fahlman[6] for a discussion).

It is this phenomenal ability to access relevant experience amidst huge tracts of memories of past experiences which penetrates all our other cognitive performances. Almost any piece of shared knowledge can be relevant to the interpretation of any anaphor in a discourse, and anaphors come thick and fast. How are these phenomenal if mundane retrieval abilities related to the encoding bottlenecks in human memory? One sketch of

an answer is that the retrieval abilities rely on the introduction of redundancy, and of a great deal of forward inferencing about the implications of new material being done at time of input. This is what places the severe limitations on speed of input and on contentfulness of the connections between materials.

DIRECT (STRUCTURAL) *VERSUS* INDIRECT (CONTENTFUL) SOLUTIONS

In von Neumann architecture computers, the binding problem is solved structurally by setting pointers from tokens of properties to a locus which identifies the individual to which they are attributed. The binding is represented regardless of the properties linked or the criteria which identify the locus. Anything can equally easily be connected to anything. This representation of binding is *direct* because something (links) actually represents bindings. These direct solutions are like binding through proper names: two predicates attributed to the same proper name are bound structurally through common reference, but, at least if the names are logically proper, there is no content to the ascription of a name beyond the identity of its denotation.

What other sorts of solutions to the binding problem are there? The most obvious contrast is with binding expressed through quantificational facts. Suppose we know that there are only two individuals in some domain, and we know the following facts:

Everything is either A or ~A.
Everything is either B or ~B.
Everything is either C or ~C.
There is an A which is B.
There is a B which is C.
There is a B which is ~C.
There is an A which is C.
There is an A and an ~A

The fact that there is an individual who is A&B&C and another that is ~A&B&~C is represented by these facts without referential terms: the property A is bound to the property ~C indirectly through quantificational facts. It is such indirect binding which will be contrasted with direct methods.

Of course, if we express this in predicate calculus, our expression of binding *within the conjunctive clauses* will depend on the apparatus of quantifiers and variables, and this apparatus is as direct and as structural as is the apparatus of referring logical constants. But if we think of how this information might be represented in people's memories we can see that, as long as we restrict the range of quantificational facts to simple existentially quantified conjunctions, they can be represented by contentful additions to the original set of facts. If we know enough properties which would serve to 'compose' the conjuncts of a statement in the domain at hand, these properties would serve to perform the necessary binding. If we know of a property H which is equivalent to the conjunction of properties A and B, then we can eliminate the apparatus of quantifiers from conjunctions of A and B. To serve the purposes of memory, we do not need a relation as precise as logical equivalence, but only one that is sufficient to pick out the combination A&B from other combinations possible in the present domain.

Suppose we have learned of a Polish bishop who is tall and sad, and a Swiss dentist who is short and sad. We have a binding problem in memory: we must discriminate our Polish bishop and Swiss dentist from possible Polish dentists and Swiss bishops. What is required is an association in general knowledge which will suffice, when added to our representation, to fix the fact that there is a Polish bishop. Here an association such as 'catholic' would do. Our general knowledge/beliefs provides this link between the two properties. We need not believe that there are no catholic Swiss dentists to see 'catholic' as a link between a religious profession and a catholic country. Needless to say, these properties recruited from general memory do not have to be neatly lexically expressible except when they are cited in papers: some ineffable feeling would serve just as well. A representation composed of existential facts provides the footholds for contentful binding of properties.

The relation between general knowledge and binding in these representations is quite different than the relation proposed by frame-based theories of memory. The associations recruited to implement binding within each existential fact do not fill our default values of variables in some overall consistent interpretation of the structure to be remembered. They are more reminiscent of the sort of material which is employed in deliberate mnemonic techniques. Talking to subjects who have done experiments similar to those described here suggests that this is just what subjects do: the sorts of examples they give of their strategies are exactly the sorts of associations that mnemonic techniques prescribe. But if they are mnemonics, they are ones which at least this group of subjects naturally employ, since the data produced is essentially uniform throughout the experimental sequence. This suggests that 'natural' memory may have much more in common with conscious mnemonic techniques than psychologists commonly acknowledge. Adult subjects at least may well have complex batteries of mnemonic methods which they are accustomed to bringing to bear, almost without noticing them, on everyday memory problems. Some of these techniques emerge in the analysis of errors below.

This similarity with prescribed mnemonic technique raises a puzzle associated with the operation of all mnemonics that require the recruitment of associations from general knowledge. This mnemonic paradox is inclined to arise as soon as such techniques are described to anyone who has never used them. Two questions immediately arise: why is it easier to remember *more* material (the target material *plus* the associations)? And how do we keep ourselves from confusing recruited material with presented material? The answer to the first question lies in the fact that the associations are already in the memory, and explains why material for which associations cannot be found will not yield to these methods. It is no good just making up random 'associations' and adding them to the presented material. The answer to the second question lies in the context of retrieval. In contexts of retrieval in which there are many constraints on possible responses there will be no danger of recruited material intruding into the retrieved material, because the context provides the only terms of description available. In a less constrained retrieval context such material might well interfere. This interference might go some way to explaining why many

Table 1. Observed and expected probabilities of occurrence of response categories

Abbreviation	*Response type*	*Observed*	*Expected*
corr	Correct	0.707	0.006
misc	Miscellaneous	0.014	0.569
sg1+	Single error on R-1 matched	0.018	0.009
sg1−	Single error on R-1 mismatched	0.025	0.015
sg2+	Single error on R-2 matched	0.024	0.009
sg2−	Single error on R-2 mismatched	0.044	0.015
ipol	Individual polarity error	0.066	0.015
is1+	Individual polarity with 'sg1+'	0.005	0.016
is1−	Individual polarity with 'sg1−'	0.004	0.023
is2+	Individual polarity with 'sg2+'	0.012	0.016
is2−	Individual polarity with 'sg2−'	0.008	0.023
2cs1	Double complementary both on R-1	0.008	0.019
2cs2	Double complementary both on R-2	0.008	0.019
2cdf	Double complementary on R-1 & R-2	0.014	0.032
dhs1	Double homogeneous on R-1	0.004	0.019
dhs2	Double homogeneous on R-1	0.007	0.019
dhdf	Double homogeneous on R-1 and R-2	0.002	0.037
ppol	Predicate polarity error	0.012	0.009
pp+s	Predicate polarity with single	0.005	0.055
mirr	Mirror image matching structure	0.008	0.049

mnemonic techniques prescribe the use of *bizarre* recruited imagery: if the material is bizarre with respect to the presented information, it can more easily be separated from it at retrieval time.

Indirect systems of representing binding allow us to explain why the ability to bind items together in memory is dependent on our general knowledge about the items to be bound. Indirect systems do this at the expense of complicating both encoding operations and the inferences necessary at the time of retrieval. How can we assess experimentally whether human memory adopts direct or indirect solutions to the binding problem?

MEMORY FOR INDIVIDUALS TASK

What is needed to study the binding problem in human memory is a task which presents material capable of many plausible combinations, each of which can be imbued with its own meaning. In this study the authors chose the task of remembering simultaneous descriptions of several individuals, and worked mostly with descriptions of pairs of people. Describing more than one individual at a time maximizes the opportunity of confusing in memory between combinations of properties. If each individual has one of two professions, nationalities, temperaments and statures, there are 136* unordered pairs of individuals defined by these four binary choices, and the task is to remember which of these combinations was presented. Although our knowledge and stereotypes of people make some combinations more likely than others, they are all quite possible.

Although a number of experiments using several methods of testing were done, the chosen method was a multiple-choice menu. A description of first one and then the other individual is picked off a menu of eight possible property values. This retrieval of information from memory is normally done immediately after reading the subject's descriptions from a micro-computer screen, sentence by sentence, and answering some questions about them. Reading is sentence by sentence self-paced because the time subjects choose to spend on each sentence is revealing of processes that go on in 'working memory', but they are not of concern here (see References 7 and 8 for more detailed discussion of both the reading times and of the error data discussed here). In a typical experiment, a subject reads about a hundred paragraphs, each describing a pair of people, in two sessions lasting perhaps an hour and a half. Each paragraph is constructed from different combinations of eight properties, but these combinations are drawn from only 48 words, 12 for each of the four categories: professions, nationalities, statures and temperaments.

As will immediately be appreciated, this is hard work. What is surprising is that people are rather good at it. In stark contrast with performance in formally similar list learning experiments (e.g. see Reference 9 for an introduction to this literature), people perform quickly and accurately, and show no build up of proactive interference, i.e. under these 'immediate' recall conditions, learning new pairs of individuals does not become either slower or less accurate as more paragraphs are learnt. This is an indication that the pairs are sufficiently distinct from each other not to interfere in the way that unstructured lists of words would do. Despite the small vocabulary, the individuals described are sufficiently richly specified to avoid the sort of devastating interference which would result if materials could not be integrated into something more than their component attributes.

Undergraduate subjects get an average of half a property wrong per paragraph on these materials. It is the pattern of errors which is revealing of the organization of the underlying representations. Each paragraph can yield 136 distinct recalls, and so it is possible to organize the error data in a 136^2 confusion matrix. This matrix is far too sparse for statistical analysis, and so the responses are classified into 20 types chosen for their theoretical interest and their frequency in the data. Table 1 shows the types of error with their observed

*There are $(((2^n)^2 - 2^n)/2) + 2^n)$ unordered pairs, where n is the number of binary property dimensions

prevalence and the proportions of possible responses which each type covers.

Errors are clearly interdependent: multiple property errors are far more frequent than they would be if an error on one property were independent of errors on other properties. Errors are more common on the second recalled individual than on the first (the data are organized here by recall order), and asymmetrical errors are more common on matched dimensions (dimensions on which both individuals have the same value). A particularly prevalent sort of response is a polarity error, in which both individuals are incorrectly recalled on one dimension. The great preponderance of these errors are individual polarity errors, in which two different values are assigned to the wrong individuals, whereas property polarity errors in which the same value is wrongly attributed to two individuals are quite uncommon. This accords with intuition: what is difficult is binding (remembering whether it was a Polish bishop and Swiss dentist or a Polish dentist and a Swiss bishop), rather than remembering property values (whether they were both Swiss or both Polish). These individual polarity errors also commonly occur with single errors on another dimension. There is considerable fine detail in the distribution of double errors that are on different dimensions, and some of these differences are usefully diagnostic of direct and indirect binding. There are very few 'miscellaneous' errors made, even though this category accounts for more than half of the total of possible responses. When things go wrong, multiple errors are more common than they would be if each of the eight properties were represented independently, yet very seldom does the whole structure fall disastrously apart. It is also noticeable that errors are more common on dimensions on which the pair of individuals mismatch (turning them into matched dimensions) than on dimensions on which they match (turning them into mismatched dimensions). This asymmetry between stimulus and response is of considerable importance in interpreting the statistical analysis and the subsequent PDP simulation.

STATISTICAL DATA MODELS

The aim of modelling the data is to show which parts of the information in these structures are represented independently, and which are dependent on parts of the representations that share common fate when they are corrupted in memory. Confusion matrix data is used both in perceptual tasks and in memory to explore psychological similarity metrics (see References 10 and 11 for two classical examples). The more psychologically similar two stimuli are, the more likely they are to be confused with each other. If we have data on how often pairs of stimuli are confused, then we can use the data to construct a similarity metric.

In order to isolate the elements of psychological similarity, the presented and recalled pairs of people are factored into propositions, and which propositions maintain and which change their truth values through the transformations observed between stimulus and response are observed. The aim is to find a set of propositions (refered to as 'features') which, when assigned weightings, can explain the observed frequencies of categories of error. Adding a feature which changes truth value with a certain error will make that error less likely: adding a feature which preserves its truth value with particular error will make that error more likely. Linear regression is used to fit weighting coefficients to sets of features with the log of the adjusted frequency of each error category as the dependent variable. The adjustment is made by dividing the observed frequency by the number of opportunities there are for each error category to occur.

Table 2. Summary of indirect feature regression model (R^2 = 0.86 deg. of fdm. = 15/100)

Feature	*Coefficient*	*Standard error*
Intercept	−4.72	–
∃x(~Bx&~Dx)	0.23	0.11
∃x(~Ax&~Dx)	0.26	0.11
∃x(~Ax&~C)	0.38	0.10
∃x(~Bx&Cx&Dx)	0.43	0.12
∃x(Ax&Bx)	0.47	0.12
∃x(~Bx&~Cx&~Dx)	0.58	0.12
∃x(Ax&Bx&~Dx)	0.25	0.10
∃x(Ax&Cx&~Dx)	0.34	0.10
∃x(~Ax&Bx~Cx)	0.43	0.11
∃x(~Ax&Bx&~Cx& Dx)	0.50	0.12
∃x(Cx)	0.68	0.09
∃x∃y(Ax&~Ay&x≠y)	0.90	0.15
∃x∃y(Bx&~By&x≠y)	0.21	0.09
∃x∃y(Dx&~Dy&x≠y)	0.73	0.07
nmat	0.22	0.07

Stenning, Shepherd and Levy[8] show how a simple model containing direct referential features, along with 'matching features' which express quantificational facts about property dimensions, can achieve a good fit to the data. This model shows that there is no tendency for all the properties of an individual to be bound to a common textually explicit property (such as the profession). The referential terms in the successful equation are identified with implicit contextual features (e.g. 'the first introduced individual'). Stenning, Patel and Levy[8] analyse data from a further experiment, and show that the referential features can be replaced by indirect quantificational features with no loss of fit to the data. Table 2 shows a re-analysis of their error data.

Using retrieval cues to fix the order of recall, they show that much of the order of recall effect is due to interference with the recall of the second individual by recall of the first, and is therefore due to a process occuring at retrieval time rather than to asymmetries in the representations. Indirect models cannot account for representational asymmetries, since any feature can equally be true of either individual.

Indirect models are, however, able to capture aspects of the data which direct models cannot. They can capture the fact that individual polarity errors are more common than property polarity errors, and some fine detail in the incidence of single errors on different dimensions. As mentioned above, errors are more common on dimensions which are mismatched. This means that stimulus and response are not treated symmetrically. No statistical model of this type can account for this fact since the equations amount to symmetrical similarity

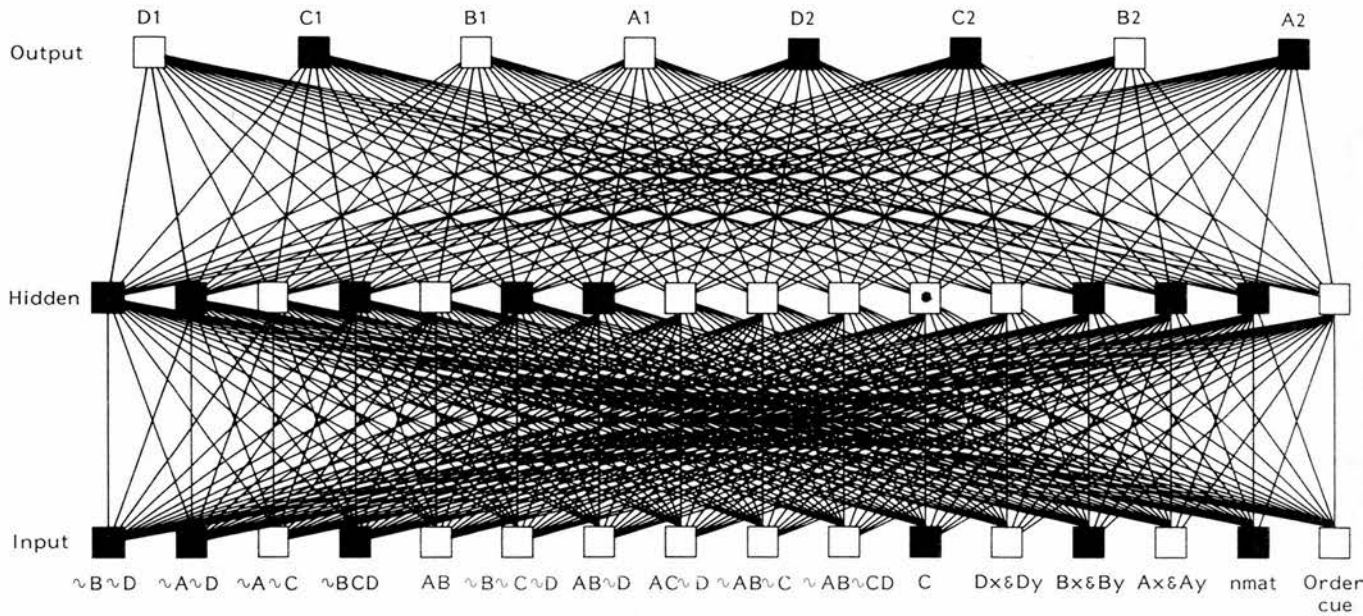


Figure 1. PDP simulation network. ■*: fully on;* □*: fully off. A = profession, B = nationality, C = temperament, D = stature*

metrics. An explanation of this asymmetry must be sought in the processes which work over these representations.

PDP SIMULATION OF INFERENCES FROM REPRESENTATION TO RESPONSE

Indirect representations of binding purchase a freedom of format and redundancy at the expense of complicating both encoding processes and inferences required at retrieval time. As shown in Table 2, it is a non-trivial inference from a set of feature values to a description of a pair of people, and the inferences are still more complicated when noise has corrupted the representations. Indirect representations are only a plausible basis for memory if they are coupled to suitable inferential mechanisms for retrieval. PDP is a natural framework for modelling memory phenomena, especially when access to those phenomena is through errors. PDP is all about interference between similar patterns. In fact, PDP systems are closely related to the regression equations which we derive directly from the data, though interesting PDP systems are non-linear.

The PDP framework could be used for modelling the process whereby error arises through the interference of similar items, but the focus here is on simulating the inferential processes involved in retrieval. A simulation is sought which can make inferences from well-formed states of the feature set to the correct output, as defined by the eight properties of the pair of individuals, and will make errors similar to those observed in human performance when noise is introduced, which leads to ill-formed inconsistent feature-value combinations.

As we know the logical relations between the features of the representation, we know a net will require hidden units to compute the required function: the problem is a complex exclusive disjunction problem, and so a feed-forward network learning by back-propagation of errors is used (see the discussion of the 'XOR' problem in Reference 12). Accordingly, a three layer net was used (Figure 1), in which the input units corresponded to the features of the regression model, taking on activations ranging from 0 (false) to 1 (true).

These input units are completely connected to a layer of 16 hidden units, which in turn are connected to two sets of four output units corresponding to the two individuals' properties. For the network to solve the binding problem it must be able to express the fact that the individuals can be recalled in either order, i.e. that either individual can be expressed on either set of output units. In order to define a function for the net to compute, there must be some input indicating which order of output is required, and so we add an extra input bit labelled 'cue'.

The network learns the function by cycling through a training set of all the unique well-formed input vectors, each paired with its correct output (240 vectors in all). The input vector corresponding to each possible pair of people is presented with an output in one order with its 'cue' unit 'off', and with an output in the other order with its 'cue' unit 'on'. The generalization that concerns us here is to behaviour when ill-formed inconsistent input vectors are presented, rather than from sub-sets of the well-formed vectors to the rest.

The back-propagation algorithm adjusts the weights of the connections in the network by comparing the activation taken on with the desired activation level. The decrease of error through the cycles of training are shown in Figure 2. The mature network performs the inference from well-formed input to output description with errors on a very few patterns.

It might be supposed that it would be difficult to learn the cueing; that is, to force the network to learn two radically different outputs for two inputs which differ by only one cue unit's level. Our evidence is that learning the position invariance of the output adds little to the difficulty of learning: the task of learning a single fixed output sequence is only slightly faster (250 epochs of training instead of 300), and the error function has a similar shape. The ability to learn this invariance easily, on the basis of these representations, is important. It

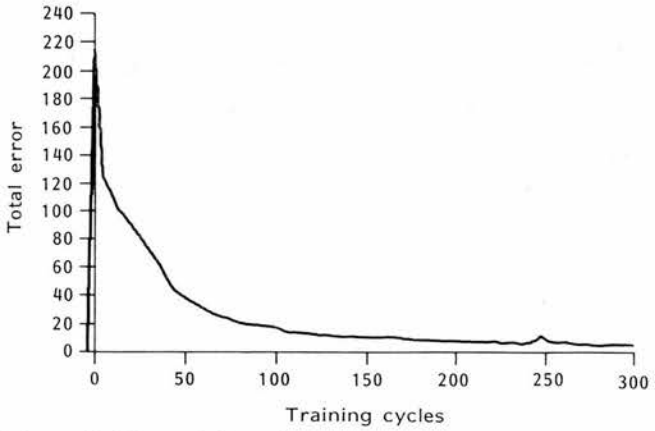


Figure 2. Network learning curve

is this invariance which demonstrates that the network has solved the binding problem. The input vector is not treated as an indissoluble gestalt, but as corresponding to a pair of sub-patterns which may be displayed in either position.

Finally, the paragraphs which the human subjects actually saw were coded into their feature representations and then subjected to random noise, i.e. each input bit's value was changed from 0 to 1 or 1 to 0 with a fixed probability. The errors in Figure 3 were produced by 3% of noise, and the figure also shows comparisons with the human memory data, the regression model on which the network is based, and the pattern produced by injecting 5% of noise directly into the eight properties of the descriptions.

The errors generated by the network obviously do not correspond to the human data as closely as the regression model's predictions: that is hardly surprising since the simulation is parameter free. The only information the network has is the features and their patterns of correlation of truth values. It contains no information about the coefficients assigned in the regression equation, nor the actual accuracies of recall of the features. But comparison of the behaviour of the network with human errors shows some striking similarities. The individual polarity plus singleton errors are particularly diagnostic. Because these errors involve three properties being wrong, they occur with extremely low frequency in the comparison simulation which applies 5% of noise directly to the eight properties. The PDP network actually overestimates the frequency of these triple errors (70 simulated > 45 observed).

The simulation captures the observation that individual polarity errors are more common than property polarity errors (59 individual > 4 property): in fact, again, it rather exceeds the effect observed, but neither category is as common as in the human memory data (101 individual > 18 property).

The network successfully captures asymmetries between stimulus and response. In the memory data, 90 responses turned matched dimension into a mismatched one, and 125 did the opposite. In the PDP simulation data, 85 responses turned a matched dimension into a mismatched one, and 179 responses did the opposite. Again, the simulation appears to accentuate an effect.

How does the network capture this effect? In a well-formed representation containing only existential terms, two individuals who are both F will ensure that all the features containing ~F are false (the corresponding input units have zero activation). Changing the activation from 0 to 1 or from 1 to 0 on *any* of these input units will produce some evidence that the dimension mismatches. There are roughly equal numbers of positive and negative occurrences of each property and its negation in the feature set, so there will generally be considerable opportunity for evidence of mismatching to arise from a matched representation when corruption occurs. In this feature set, a representation of a mismatched dimension requires that some F features and some ~F features will be true, and the probability that this situation will change (i.e. *all* the F features or *all* the ~F features will become false) during corruption is correspondingly slight. The natural property of this type of representation captures the observed phenomenon in a parameterless way.

What this network without referential terms cannot do is capture the difference in error rate between the first and second recalled individuals. Since all of the features will be true of each of the individuals equally as often, no structural asymmetry can be represented. However, as mentioned above, most of the observed difference in accuracy between the individuals is due to retrieval processes which work over the representations. Representations which did incorporate these differences could not account for the cueing effects observed in other experiments[7]. Further work will be required to simulate the process of interference during recall.

CONCLUSION

It has been argued that the binding of properties to individuals in this task is accomplished by diffuse memory representations that only contain information about which of some fragmentary combinations of properties are instantiated in a particular domain. These diffuse representations are required to account for many phenomena observed in the patterning of errors observed. They must be coupled with processes for performing the complex inferences required from states of representation to best-fitting response. PDP architectures offer a direct way of modelling these inferences, and can reproduce some of the important error patterns without the adjustment of any tuning parameters. The authors propose to extend this research in the direction of showing that PDP systems can also model the contentful binding going on within the features in the models presented. Such a model would remove the requirement for modelling corruption by the introduction of random noise.

To computational eyes, this solution to binding can appear *ad hoc*. How does it relate to other situations in which there are more individuals, more properties, more relations and so forth? It is obvious that at least some different features will have to be used. Will we not end up with a different system for every situation? To some extent the answer may well be 'yes', but people do have a large battery of rather particular mnemonic devices which they bring to bear in different situations. For example, when we present them triples of individuals they start using 'odd-man-out' coding. What is important is not whether these features are general, but whether simple fragmentary quantificational information of this

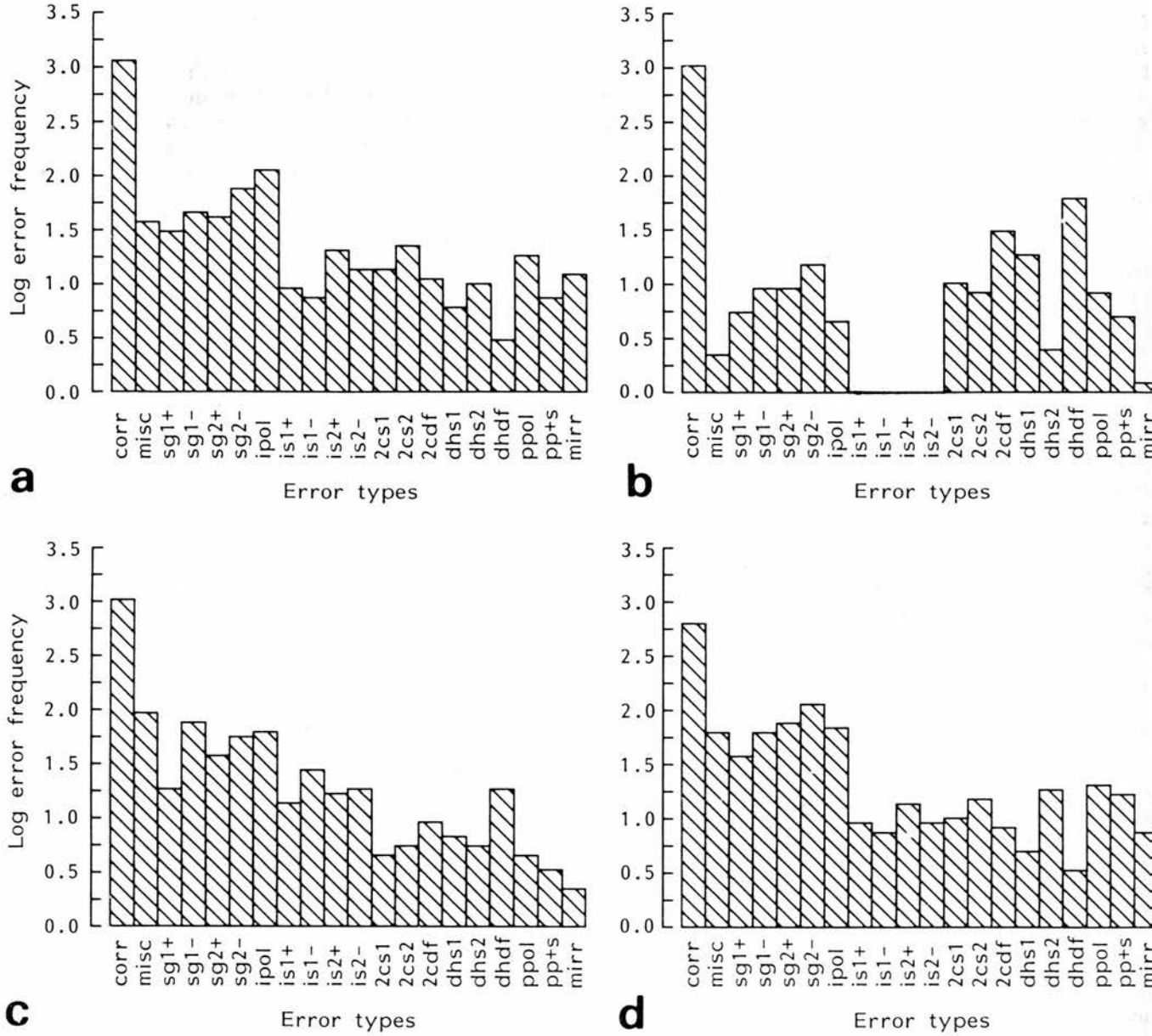


Figure 3. Error frequencies for (a) human memory data, (b) 5% noise applied directly to descriptions of individuals, (c) PDP network simulation — 3% noise, and (d) regression model predictions

type can be used to solve the binding problem *in the range of situations in which people can solve the problem.* This latter qualification is crucial, and has often caused gulfs of misunderstanding between computer scientists and psychologists. What is wrong with simple direct von Neumann solutions to the binding problem *as theories of mental representation* is that they cannot explain why people are sensitive to content, and do not have indefinitely large working memories.

So, the crucial question is, does this sort of solution scale appropriately with human abilities? Much more experimentation will be required to find out what human abilities are in related tasks, and much more analysis is required to show how these representations scale with increased numbers of properties, individuals and fragment sizes. At present we can only say that the indications are that increasing numbers of property dimensions seems to cause people less trouble than increasing numbers of individuals. People can cope almost as well with pairs of people with six binary properties. We have little firm evidence about large numbers of individuals save to say they seem to be harder. This may also be so with the properties of these representational systems. Increasing the number of individuals increases the complexity of the inferences from representation to response more than increasing numbers of dimensions.

In assessing the capacity of human memory in these tasks we are, of course, asking about simultaneous learning capacity rather than ultimate capacity of long-term memory. Clearly people can distinguish large numbers of individuals in their total memories. However, such memories are sub-divided into sub-domains, and are built up over long periods of time with rich distinguishing information. Very little is known about the time course of the memories which are used in the Memory for Individuals Task (MIT). Some residue certainly remains: tested six weeks later, subjects could correctly pick out pairs of individuals from the 100 or so they had previously seen from ones which they had not about

65% of the time. The foils are constructed from the same small vocabulary set of descriptors, and so are highly confusable. Little is known about what type of information survives, and whether it changes as time goes by, but these memories are extremely durable by the standards of rote memory tasks.

How does this laboratory task relate to real-world tasks, and what morals does it suggest for the designer? The binding problem is an extremely general knowledge representation problem. If we assume it poses problems for people wherever many assignments of properties to individuals yield individuals which are 'similar' as far as general knowledge allows, we have quite a good prescription for finding places where people experience memory as a problem. When we park our car in the same car park every day, we are left with a large number of very similar 'parking occasions' with many more 'possible but not actual' occasions, and little general context for knowledge rich solutions to differentiate them. When we are editing a series of similar files all containing similar programs and with rather arbitrary names, or tracing the values of a set of variables through a program where they represent rather similar information, we are prone to experience memory overload, and our likelihood of success is related to our ability to differentiate the combinations by associating them through what may be extraneous general knowledge. If our theory about binding is correct, adults are rather good at identifying where problems will arise, at recruiting all sorts of heterogeneous information about interrelations between bits of the structures involved, and at using this to differentiate assignments in a rich context of retrieval. Memory loads will be very different for the novice than for those with an extensive knowledge of the range of possibilities in a situation, and of the 'meaning' of the different combinations of properties that define them.

From this work it is possible to distinguish between the theory of representations which is extracted by regression modelling of error data, and the particular implementation of inference in the PDP network. The two are closely related, but not inseparable. Because the logic of the features is explicit, the inferences could be implemented in any standard theorem prover. A mechanism would then have to be added to account for how the inconsistencies arising in redundant representations subject to noise are resolved, and a best approximation to the right conclusion found. Whether this could be done in any principled way which generated similar errors is an interesting question. Pursuit of alternative implementations would throw light on what is particular to each of them. The underlying attraction of PDP systems to a psychologist wanting to model human information processing is the hope that its strengths and limitations coincide better with human abilities.

In view of the popularity of the topic, it may be worth clarifying the role of PDP simulation in this work. First, the work does not focus on learning: the back-propagation algorithm provides a convenient 'automatic programming tool' for setting up a network which will perform the desired inferences, but in principle it would serve the same function if it were hard wired. That is because we have no evidence about how our subjects learnt the encoding retrieval techniques which they use. This experiment was about remembering individuals, not learning the logic that underlies the inferences involved — subjects learnt that long before they became subjects. The automatic learning makes our argument more powerful, because it shows that no 'tweaking' has gone on to adjust the error behaviour, other than the back-propagation learning of the logical relations between features. The data is fed straight into the regression, and the features in the regression model are then incorporated directly into the network.

Second, the PDP approach has been much criticised by computer scientists, because in order to compute some simple functions (parity, symmetry, connectedness, etc.) it can be shown that the resources required in a network (number of connections, cycles of training, etc.) grow dramatically with the size of the problem, in a combinatorial explosion. The *locus classicus* of this argument is Minsky and Papert[13]. However, people are also generally inept at computing these general functions as the size of the problem increases: biology is not computer science. What would be really damaging would be to show that in areas where people do not fall down with the size of the problem (in such areas as storage capacity and speed of retrieval), PDP systems cannot be constructed without explosive resource implications. The state-of-the-art in engineering PDP systems is primitive, and so is our detailed knowledge of peoples' abilities in controlled situations such as the experiment presented above. Theorems about the size of a system necessary to compute connectedness way beyond human capacity are not going to help us evaluate these systems as theories of human computational architecture. This work is a demonstration that, what to a computer scientist is a 'toy system', is capable of modelling at least one aspect of human ability in a manageable way.

Third, PDP systems are not merely relevant to 'low level implementational' issues, and neither is the organization of memory such a low-level resource issue. The content sensitivity of human memory is a high level cognitive phenomenon: one of the issues on which cognitive science was established. Human cognition is equipped with a retrieval system which can mobilize the relevant information from masses of stored data quickly on the basis of unreliable search cues. This resource enables all other skills — language, mathematics, chess, vision — to be mounted on machinery which is extremely slow. It also gives mental life its phenomenal quality. Far from being an implementational detail which can be left until we get the right account of these abilities, memory is the common factor that holds clues to how all of these are possible.

Fourth, PDP modelling is not the search for some Baconian discovery procedure which will solve the induction problem: networks are tools for doing cognitive science, not cognitive scientists. Understanding human memory is not a matter of setting some enormous network going and watching it learn to be a human being. What is required is a careful analysis of what people do, and careful analysis of what aspects of network architecture can simulate the important phenomena. That is why the authors' preference is for analysing material where there is a clear logical model of the input and the response. Much more work is taking place applying networks to perceptual problems where the nature of the input is ill-understood, and from an

engineering point of view there are good reasons why this should be.

Occasionally, it is claimed that it is no easier to understand how a PDP system works than to understand how the simulated human being performs the task in the first place: to this a psychologist can only propose a spell in the laboratory. Weight tables may be complicated, but there are many statistical techniques for analysing them. It is not assumed that there is nothing of a generality above the level of individual units' behaviour to be said about how networks work. To analyse how a network achieves position independent output is one of the things we intend to do next.

ACKNOWLEDGEMENTS

This research was supported by SERC (Alvey) Grant #GR/D10138 and by J S McDonnell Foundation Grant #549210.

REFERENCES

1 **Hirst, G** *Anaphors in natural language understanding: a survey* Springer-Verlag, USA (1981) pp 128
2 **Tulving, E** 'Episodic and semantic memory' in *Organization of Memory* Academic Press, USA (1972)
3 **Bartlett, F C** *Remembering: a study in experimental and social psychology* Cambridge University Press, UK (1932)
4 **Schank, R C and Abelson, R P** 'Scripts, plans and knowledge' in **Johnson-Laird, P N and Wason, P C (eds)** *Thinking* Cambridge University Press, UK (1977)
5 **Ebbinghaus, H** *Uber das Gedachtnis* Duncker and Humblot, Austria (1855)
6 **Fahlman, S E** *Representing Implicit Knowledge* Lawrence Erlbaum Associates, USA (1981)
7 **Stenning, K, Patel, M J and Levy, J** *The 'Binding Problem' in human memory: some effects of referential discontinuity on the construction of representations for individuals* Technical Report, University of Edinburgh, UK (1987)
8 **Stenning, K, Shepherd, M and Levy, J** 'On the construction of representations for individuals during text comprehension' *Research Paper No 9* Centre for Cognitive Science, University of Edinburgh, UK (1987)
9 **Baddeley, A D** *The Psychology of Memory* Basic Books, USA (1976)
10 **Conrad, R** 'Acoustic confusion in immediate memory' *British J. Psychol.* Vol 55 (1964) pp 75–84
11 **Miller, G A and Nicely, P** 'An analysis of perceptual confusions among some English consonants' *J. of the Acoustical Soc. of America* Vol 27 (1955) pp 338–352
12 **McClelland, J L and Rumelhart, D E (eds)** *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* MIT Press, USA (1986)
13 **Minsky, M and Papert, S** *Perceptions: An Introduction to Computational Geometry* MIT Press, USA (1968)